Downloaded from orbit.dtu.dk on: May 05, 2023

DTU Library

Design of Computer Experiments

Dehlendorff, Christian

Publication date:
2010

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Dehlendorff, C. (2010). *Design of Computer Experiments*. Technical University of Denmark. IMM-PHD-2010-237

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Design of Computer Experiments

Christian Dehlendorff

Kongens Lyngby 2010
IMM-PHD-2010-237

Technical University of Denmark
Informatics and Mathematical Modelling
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-PHD: ISSN

Summary

The main topic of this thesis is design and analysis of computer and simulation experiments and is dealt with in six papers and a summary report.

Simulation and computer models have in recent years received increasingly more attention due to their increasing complexity and usability. Software packages make the development of rather complicated computer models using predefined building blocks possible. This implies that the range of phenomenas that are analyzed by means of a computer model has expanded significantly. As the complexity grows so does the need for efficient experimental designs and analysis methods, since the complex computer models often are expensive to use in terms of computer time.

The choice of performance parameter is an important part of the analysis of computer and simulation models and Paper A introduces a new statistic for waiting times in health care units. The statistic is a measure of the extent of long waiting times, which are known both to be the most bothersome and to have the greatest impact on patient satisfaction. A simulation model for an orthopedic surgical unit at a hospital illustrates the benefits of using the measure.

Another important consideration in connection to simulation models is the design of experiments, which is the decision of which of the possible configurations of the simulation model that should be tested. Since the possible configurations are numerous and the time to test a single configuration may take minutes or hours of computer time, the number of configurations that can be tested is limited. Papers B and C introduce a novel experimental plan for simulation models

having two types of input factors. The plan differentiates between factors that can be controlled in both the simulation model and the physical system and factors that are only controllable in the simulation model but simply observed in the physical system. Factors that only are controllable in the simulation model are called uncontrollable factors and they correspond to the environmental factors influencing the physical system. Applying the experimental framework on the simulation model in Paper A shows that the effects of changes in the uncontrollable factors are better understood with the proposed design compared to the alternative and commonly used methods.

In papers D and E a modeling framework for analyzing simulation models with multiple noise sources is presented. It is shown that the sources of variation of the simulation model can be divided in two components corresponding to changes in the environmental factors (the uncontrollable factor settings) and to random variation. Moreover, the structure of the environmental effects can be estimated, which can be used to put the system in a more robust operating mode.

The interpolation technique called Kriging is the topic of Paper F, which is a widely applied technique for building so called models-for-the-model (metamodels). We propose a method that handles both qualitative and quantitative factors, which is not covered by the standard model. Fitting the final Kriging model is done in two stages each based on fitting regular Kriging models. It is shown that this method works well on a realistic example such as a simulation model for a surgical unit.

Resumé

Hovedområderne i denne afhandling er design and analyse af computer- og simulationseksperimenter. De er afdækket i seks artikler samt en sammenfattende introduktion.

Simulations- og computereksperimenter har i de senere år fået stadig større bevågenhed på grund af kompleksiteten og anvendeligheden af disse modeller. Der findes adskillelige software pakker, der muliggør udvikling af meget komplekse modeller ved hjælp af prædefinerede byggeblokke. Dette betyder, at stadig flere systemer kan analyseres ved hjælp af computermodeller. Med den øgede kompleksitet er behovet for effektive eksperimentelle planer og analyse metoder steget, idet de komplekse modeller typisk er tidskrævende at bruge.

Valg af performance parameter er en vigtig del af analysen af computer- og simulationsmodeller, og i artikel A introduceres en ny statistik for ventetider i hospitalsenheder. Statistikken er et mål for størrelsen og udbredelsen af lange ventetider, som er de mest generende og har den største indflydelse på patienttilfredsheden. En simulationsmodel for en ortopædkirurgisk operationsgang på et hospital blev brugt til at illustrere fordelene ved statistikken.

En vigtig overvejelse i forbindelse med simulationsmodeller er den eksperimentelle plan, hvilket er valget af hvilke af de mulige konfigurationer af simulationsmodellen, der skal afprøves. De mulige konfigurationer for en simulationsmodel er ofte mange, og tiden for at teste en enkelt konfiguration kan tage flere minutter eller timer i computertid. Dette betyder, at antallet af konfigurationer, der kan testes, er begrænset. Artiklerne B og C introducerer en ny eksperimentel plan for simulationsmodeller, der har to typer af input faktorer. Planen skelner mellem faktorer, der kan kontrolleres i modellen og i det fysiske sys-

tem, og faktorer, der kun kan kontrolleres i modellen. Sidstnævnte kaldes også ukontrollerbare faktorer og svarer til de miljøfaktorer, der influerer det fysiske system. For simulationsmodellen for den kirurgiske operationsgang blev det vist, at sammenlignet med eksisterende eksperimentelle planer giver det nye design en bedre forståelse af de ukontrollerbare faktorers betydning.

I artikel D og E blev et framework til analyse af simulationsmodeller med flere støjkilder præsenteret. Det blev vist, at variationskilderne kan opdeles i to komponenter svarende til ændringer i de ukontrollerbare faktorer og tilfældig variation. Ydermere blev det vist, at effekten af variationer i de ukontrollerbare faktorer kan estimeres, hvilket kan udnyttes til at sætte systemet i en mere robust konfiguration.

Artikel F omhandler interpolationsteknikken Kriging, som er en ofte anvendt teknik til at estimere såkaldte modeller for modellen (meta-modeller). En ny metode, der muliggør Kriging for simulationmodeller med både kvalitative og kvantitative faktorer, introduceres. Krigingmodellen estimeres i to skridt, som begge består af estimation af sædvanlige Krigingmodeller. Metoden testes på simulationsmodellen for den kirurgiske operationsgang, hvor det vises, at metoden virker bedre end eksisterende metoder.

Preface

This thesis was prepared at DTU Informatics (Informatics and Mathematical Modelling) at the Technical University of Denmark in partial fulfillment of the requirements for acquiring the Ph.D. degree in engineering. It was funded by the Technical University of Denmark and was supervised by Klaus Kaae Andersen and Murat Kulahci.

The thesis deals with different aspects of design and analysis of computer and simulation experiments. The thesis consists of a summary report and a collection of six research papers written during the period 2007–2010, and elsewhere published.

Lyngby, August 2010

Christian Dehlendorff

Papers included in the thesis

A Christian Dehlendorff, Murat Kulahci, Søren Merser and Klaus Kaae Andersen, *Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units.* Published in Quality Technology and Quantitative Management (2009). N C T U Press,. Vol. 7(3), p. 321-336

B Christian Dehlendorff, Murat Kulahci and Klaus Kaae Andersen, *Designing Simulation Experiments with Controllable and Uncontrollable Factors.* Published in Proceedings of Proceedings of the 2008 Winter Simulation Conference, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

C Christian Dehlendorff, Murat Kulahci and Klaus Kaae Andersen, *Designing simulation experiments with controllable and uncontrollable factors for applications in health care.* Published in Journal of the Royal Statistical Society, series C (2011), 1

D Christian Dehlendorff, Murat Kulahci and Klaus Kaae Andersen, *Analysis of Computer Experiments with Multiple Noise Sources (European Network for Business and Industrial Statistics).* Published in Proceedings of ENBIS8, Athens 2008, non peer-reviewed

E Christian Dehlendorff, Murat Kulahci and Klaus Kaae Andersen, *Analysis of Computer Experiments with Multiple Noise Sources.* Published in Quality and Reliability Engineering International, Volume 26 Issue 2, March 2010, p. 147-155 (Special issue for ENBIS8)

F Christian Dehlendorff, Murat Kulahci and Klaus Kaae Andersen, *2-stage approach for Kriging for simulation experiments with quantitative and qualitative factors.* Submitted to Technometrics

Acknowledgements

First of all I would like to thank my two supervisors Klaus Kaae Andersen and Murat Kulahci for all their valuable comments, ideas, suggestions and encouragements.

I would also like to thank Dr. John Fowler and Dr. Douglas Montgomery for an interesting stay at Arizona State University. And Murat and his wife Stina for all their help during my stay in Arizona.

The staff at the orthopedic surgical unit at Gentofte Hospital was helpful in the collection of the data for the simulation model. Michel Boeckstyns assisted in the description of the surgical unit and collecting data. Søren Merser from Frederiksberg Hospital has been a great help in building the simulation model and providing the contact to Gentofte Hospital.

Also Klaus Kaae Andersen and Henrik Spliid are to be thanked for the many interesting projects that I have had the possibility to participate in during my employments at IMM's Statistical Consultancy Center.

During my ph.d. study I have had the great pleasure of working together with several other researchers in areas outside the topic of my thesis. This has been extremely interesting and useful, so thank you to Sigrid Tibæk, Tom Skyhøj Olsen and Rigmor Jensen.

A special thanks to my wonderful wife Maiken, who has supported me all the way and listened patiently to my latest findings, results and challenges. Without her the last three years would definitely not have been as joyful and good. Finally,

a thank you to my daughter Isabella for keeping my spirits up with her cute little smiles and always positive ”dada”s during the last eleven months.

Contents

Chapter 1

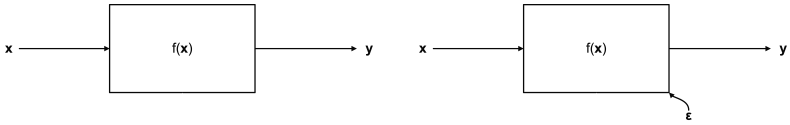
Introduction

The title of this thesis is "design of computer experiments" and it deals with the planning and analysis of experiments with a computer model as a replacement for physical experimentation. Computer models are used in many areas in which physical experimentation is either not possible or expensive. One example of a physical system in which experimentation is impossible (or at least very limited) is an orthopedic surgical unit at a hospital. For such a system, patient safety concerns restrict the experimentation and moreover the cost of certain experiments may make them infeasible to do, e.g., putting in an extra operating room to test how it would improve the performance is a very expensive experiment. Another example is crash testing of cars, which can be simulated with a computer model in order to save the costs of actually crashing a car. Using a computer model allows the designers and engineers to test many configurations at a low cost. A third example is the design of hip replacements (Chang et al., 1999), which may reduce the costs for clinical trials significantly.

1.1 Simulation models

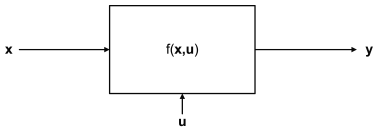
A computer model generates a set of outputs (although usually only one outcome is considered at a time) that depends on a set of input factors. For a surgical unit the input factors are, e.g., the number of doctors and operating rooms,

whereas the output, e.g., is the patient waiting time. Computer models are usually classified as being either deterministic or stochastic; that is, the output either stays the same (deterministic) or varies (stochastic) for replicated runs with the same settings of the input factors.

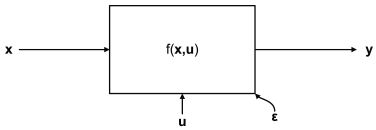


(a) Computer model with one factor type

(b) Computer model with one factor type and stochastic output



(c) Computer model with two factor types



(d) Computer model with two factor types and stochastic output

Figure 1.1: Basic structures for computer models

Four basic structures of computer models are shown in Figure 1.1. The most simple model (Figure 1.1(a)) is a model which takes an input vector, x, corresponding to several variables and generates the output, y. The output may also be influenced by a stochastic component as indicated by ϵ in Figure 1.1(b), e.g., the arrival times of acute patients at the surgical unit. Another disturbance is environmental/uncontrollable factors such as the arrival rate of acute patients at a surgical unit, which is indicated by the input u in Figures 1.1(c) and 1.1(d). The uncontrollable factors may significantly influence the output, which implies that the signal, $f(x, u)$, becomes a function of both the controllable input factors, x, and the uncontrollable input factors, u. Likewise the stochastic component may influence the output from one run to the next for the stochastic computer model.

A subtype of computer models is simulation models and in this thesis a discrete event simulation model is considered. In such a model a series of events is simulated using a computer. The case study in this thesis is a model for an orthopedic surgical unit at a hospital, which simulates the patients' route from the ward (or the emergency room) to the discharge. Animation is included in the model as a tool for verifying the patient and staff flow in the model, which is a valuable tool for presenting the model as illustrated in Figure 1.2.

Several performance measures are possible outputs for the surgical unit, e.g., waiting time and patient throughput. In this thesis the performance of the unit

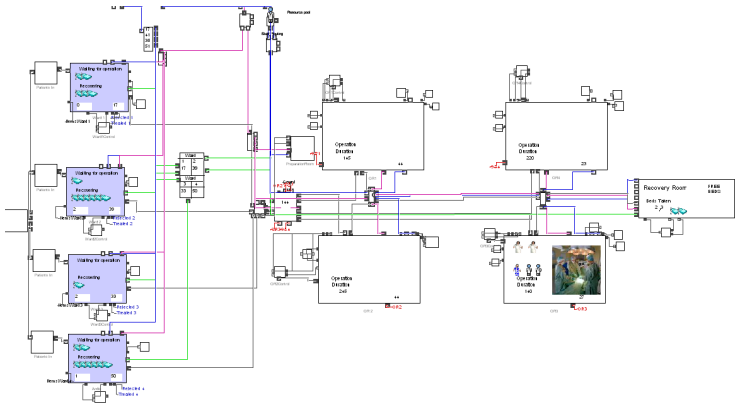


Figure 1.2: Extend model of surgical unit

is primarily measured by the extent of long waiting times since they from a patient perspective are the most bothersome. In Paper A a new measure for waiting time is introduced and compared to other existing measures. The measure is called the Conditional Value at Risk waiting time (CVaR) and measures the extent of long waiting times. In Papers C-E CVaR is reconsidered together with the number of patients treated and the fraction of planned surgery being done outside regular hours. The latter indicates the level of overtime needed. The surgical unit is used as case-study throughout the thesis and the model is described in more detail in section 2.2.

1.2 Experimental design

Computer models are often very complicated and hence may take long time to run. This implies that simply trying all possible combinations of the input factors becomes computationally infeasible, e.g., the simulation model in section 2.2 has 16 inputs and if two settings are considered for each input this gives a total simulation time of 45 days (a single run takes seven minutes to complete). Much of the literature on computer experiments is therefore related to choosing the experiments to be performed, i.e., the settings of the inputs to be tested. Such a selection of experiments is called an experimental design.

An experimental design consists of a set of experiments called design sites or runs. One such run corresponds to one specific setting of the s input factors to the model. The objective of an experimental plan is typically to choose the runs in such a way that the information in the output (and thus the model)

is maximized. In computer experiments both the costs of a single run and the number of input factor are typically high, which implies that only relatively few runs in a high dimensional space can be chosen.

The experimental plan also depends on which of the four model types in section 1.1 the computer model belongs to. For stochastic computer models replications, i.e., repeated runs of the model with the same input setting x, yields additional information of the stochastic components, whereas repetitions for deterministic computer models are redundant. The presence of uncontrollable factors as in Figures 1.1(c) and 1.1(d) also implies different experimental designs compared to the first two model types in Figures 1.1(a) and 1.1(b), since the controllable and uncontrollable factors have different interpretation in the physical system and are therefore treated differently in the design and analysis of the computer model. The design of computer experiments is discussed in more detail in Chapter 3 and a new experimental plan is proposed in Papers B and C.

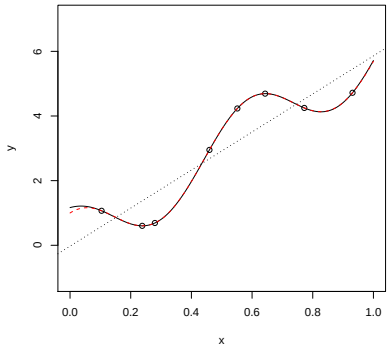
1.3 Output analysis

The second major topic of computer experiments is the analysis of the output generated from the experimental design. One objective of output analysis may be to find the optimal setting of the system, e.g., how to setup a surgical unit such that the maximum number of patients is treated. Another objective could be to build a (simpler) model for the computer model. Such a model-for-the-model is called a meta-model and is (and should be) considerable faster to run compared to the actual computer model. The computer model corresponds to an equivalent but unknown (and perhaps very complex) mathematical model and the meta-model is an approximation of this unknown model. Such a meta-model may be used for optimization in order to avoid the computational costs of using a time consuming computer model.

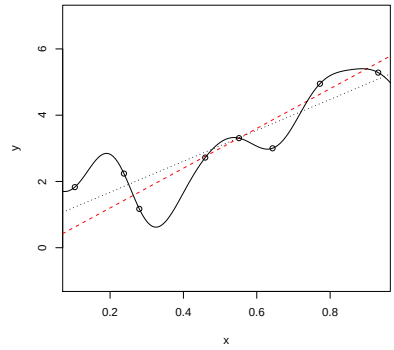
A natural question is: Why would anyone construct a complicated computer model if it can be reduced to a simpler model? Considering a surgical unit at a hospital, it may not be very clear how the relationship between the number of different staff types and the patient waiting time is. However, modeling the processes and resources needed for each sub-process is more intuitive and interpretable. The complex model may then be a result of combining several simpler models of sub-processes. Thus, modeling the quantity of interest indirectly may sometimes be the only feasible approach.

The methods used in the output analysis depend on the type of the computer

model, i.e., whether the output is deterministic or stochastic. In the deterministic case a natural criterion is that the model for the output interpolates the data; that is, the meta-model equals the model output at the design sites. Figure 1.3(a) shows a meta-model for a deterministic computer model. It can be seen that the meta-model (an interpolator called Kriging) is an adequate description of the underlying signal, whereas the linear regression line ignores the periodic part of the underlying model. From Figure 1.3(b) it can be seen that interpolating the output from a stochastic computer model gives a highly wiggly and inappropriate predictor, whereas the regression line is seen to be a better description of the underlying model. In the stochastic setting a vast literature from the analysis of physical experimentation exists, which also (potentially with some modifications) can be applied for computer models.



(a) Deterministic output with underlying model given as: $y = \cos(6.8\pi x/2) + 6x$



(b) Stochastic output with underlying model given as $y = 6x + \epsilon$

Figure 1.3: Examples of deterministic (a) and stochastic output (b), where "o" is the observations, the solid black lines are Kriging interpolators (see section 4.1), the red dashed lines are the true signals and the black dotted lines are linear regression lines (see section 4.2)

1.4 Outline of the thesis

This thesis consists of three major topics, simulation, design of experiments and output analysis as outlined in this chapter. In Chapter 2 a general introduction to simulation is given followed by an introduction to experimental design in Chapter 3. Moreover, a case-study is introduced in section 2.2 and used

throughout as motivating example. In Chapter 4 an introduction to the different analysis methods is given, which includes both regression and interpolation techniques. The included papers in Appendix A-F are summarized in Chapter 5 and the main conclusions given in Chapter 6.

Chapter 2

Simulation models

The literature concerning the design and analysis of deterministic simulation models is usually covered by the name: "Design and Analysis of Computer Experiments" (DACE) and is described by for example Sacks et al. (1989b). In the book by Kleijnen (2008) design and analysis of simulation experiments (DASE) are presented for both deterministic and stochastic simulation. A simulation model is an example of a computer model and can be either deterministic or stochastic. In this thesis a simulation model is used as case-study and it is described in more detail in section 2.2.

2.1 Model types

Simulation models are as for computer models divided into two classes: deterministic and stochastic. These two classes of simulation models are different both in terms of the type of physical phenomena they model, the experimental designs to apply and the analysis methods to use. In this chapter we briefly introduce simulation and the case-study, whereas design and analysis of simulation experiments are covered in Chapters 3 and 4, respectively.

In deterministic simulation the simulation model generates the same output for replicated runs with the same settings of the input factors. Kleijnen (2008) gives

several examples of deterministic simulation models including the "IMAGE" model for the increasing global temperatures (Bettonvil and Kleijnen, 1997). Deterministic simulation models behave differently from physical phenomena since repeated runs with the same settings yield exactly the same output. In physical experiments all factors can usually not be controlled completely and hence the outcome changes from one replicate to the next. This implies that different experimental designs and analysis techniques are needed for deterministic simulation models (Sacks et al., 1989a, Fang et al., 2006).

Many simulation models however involve some sort of stochastic disturbance making the output also stochastic and thus repeated runs with the same input give different output. The stochastic components are procedures, arrival processes, etc., which are generated by streams of random numbers. The stream is controlled by a seed, which is a number that initialize the state of the generator. The variation coming from the stochastic components implies that the model output behaves more like a physical experiment, i.e., the stochastic components somehow correspond to having the experimental error in physical experimentation.

Although stochastic simulation is seen to be more similar to physical experimentation in contrast to deterministic simulation, it is important to note that the variation in the output is artificially generated and controlled in the simulation model. In discrete event simulation the seed controls the stream of random numbers, which are used to generate stochastic arrival processes etc. This implies that the simulation model can be put in a deterministic operating mode by using the same seed. Controlling the seed is utilized in the variance reduction technique known as common random numbers (CRN) (Schruben and Margolin, 1978, Donohue, 1995, Banks et al., 2005, Kleijnen, 2008).

Another difference compared to physical experimentation is that environmental factors in simulation models can be controlled, i.e., the arrival rate of acute patients to a surgical unit can be controlled in the simulation model but not in the physical system. Moreover, the uncontrollable factors are required to have values assigned in each run, which implies that the settings of these factors become an important part of the experimental plan. Simulation models are as such the ideal experiment, since all sources of variation can be controlled.

An often used simulation technique is Discrete Event Simulation (DES), which is a simulation type where the system changes at discrete time points corresponding to a series of events (Law and Kelton, 2000). An event is, e.g., that a patient arrives at a hospital unit or a surgeon is called to the operating room at a surgical unit at a hospital unit such as in the case-study presented in section 2.2. The simulation model is controlled by a clock, which jumps to the time point for the next event on the event stack, performs the event, updates

the event stack, jumps to the next event and set the clock, etc.

2.2 Case-study: a surgical unit at a hospital

Within health care simulation is a widely used technique due to the limitations of physical experimentation in these systems (see for example Brailsford (2007)). Moreover, since health care budgets not only tend to be large but also increasing in size there is a potential for significant savings. The long list of applications of simulation in health care covers topics such as disease modeling, e.g., the spread of HIV (Mellor et al., 2007) and optimization of hospital units, e.g., optimizing an emergency department (Ferrin and McBroom, 2007). Another example is the simulation of pandemic influenza preparedness plans as considered by Lant et al. (2008), who evaluate different plans for evacuating a public university during a pandemic influenza using simulation. All three examples illustrate cases where physical experimentation is either impossible (Mellor et al., 2007, Lant et al., 2008) or too expensive (Ferrin and McBroom, 2007).

We consider a discrete event simulation model for an orthopedic surgical unit, which is implemented in the simulation software Extend (Krahl, 2002) and controlled from a Visual Basics for Applications (VBA) script in Excel. A single run corresponds to simulating six months operation (approximately 2000 surgical procedures) with a warm-up period of one week, which in Dehlendorff et al. (2010b) was shown to be a good compromise between simulation time and accuracy. The model takes approximately seven minutes to complete a single run, which is long enough to prohibit brute force analysis, i.e., running all possible combinations of factor settings.

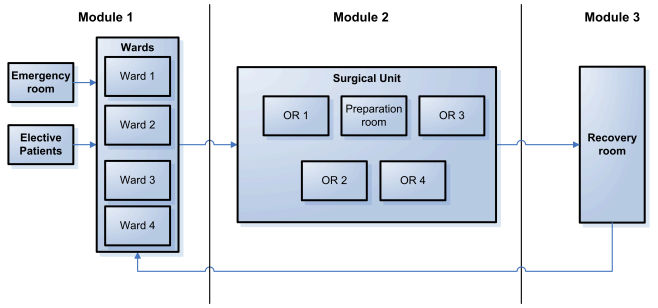


Figure 2.1: Outline of surgical unit

The outline of the surgical unit is given in Figure 2.1. It consists of three main modules: arrival, treatment and recovery. Patients arrive from either one of the

wards or from the emergency room. They are either acute or elective, i.e., an acute patient arrives from the emergency room (or from other departments in the hospital) for an operation not a planned in advance, whereas the operations for the elective patients are scheduled. In the simulation model the staff is controlled through resource pools, e.g., a pool for surgeons (as well as other staff) and a pool for operating rooms. The pools contain the idle resources and release them as soon as they become available when a procedure makes a request.

The route through the surgical unit consists of several stages as outlined in Figure 2.2. The patients arrive for either planned or acute operations and are admitted to a ward (a separate ward is reserved for the acute patients) and thereafter brought to the surgical unit. At the surgical unit the patients are sedated and prepared for surgery either in the operating room or in a preparation room and then brought to the operating room. After surgery the patients are transported to the recovery room for wake up and thereafter returned back to the ward for final recovery and discharge.

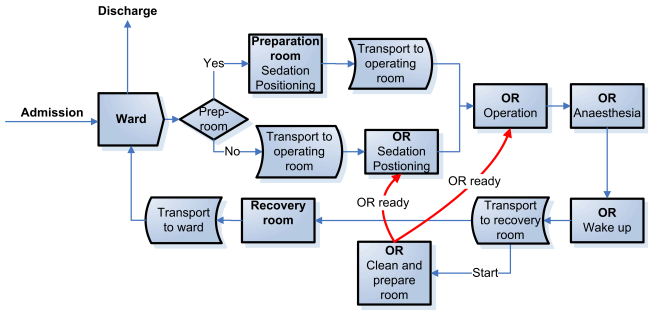


Figure 2.2: Flowchart for the patient's route through the orthopedic surgical unit

For each process in Figure 2.2, teams consisting of potentially multiple staff groups are required, e.g., for transportation of patients a porter is required, for sedation an anesthesiologist is required and for the surgical procedure nurses and surgeons are required. It entails a delay for the patient if one or more of the required resource pools are empty corresponding to the time it takes before all required resources become available.

The performance of the surgical unit may also be influenced by its surroundings, e.g., the arrival rate of acute patients can usually not be controlled in the physical system. Since the system may behave very differently depending on the settings of these uncontrollable factors, they are also included in the model. The controllable and uncontrollable factors are summarized in Table 2.1, where

a controllable factor is controllable in both the model and the physical system and an uncontrollable factor only in the model.

Type	Factors	
Controllable	*Porters*	*Anesthesiologists*
	ORs	*Recovery beds*
	Cleaning teams	*Elective patients*
	Operating days	*Acute intake*
Uncontrollable	*Porters occupied*	*Anesthesiologist occupied*
	OR cleaning time	*Recovery bed occupied*
	Cleaning teams occupied	*Surgeon occupied*
	Length of procedures	*Acute arrival rate*

Table 2.1: Factors used in simulation model for surgical unit

The performance of the surgical unit is measured by the waiting time experienced by the patients. Bielen and Demoulin (2007) show that patient satisfaction decreases as the waiting time increases; that is, from a patient satisfaction point of view long waiting times are troublesome. In Paper A a statistic, CVaR, for measuring the extent of long waiting time is introduced, which is used as primary outcome in the remainder of the thesis. Figure 2.3 shows two waiting time distributions: the gamma distributions $\Gamma(2,1)$ and $\Gamma(10,5)$. The expected waiting time is for both distributions two time units, but the lengths of the tails are very different. The focus in this thesis is the extent of long waiting time and CVaR, which is marked with vertical lines in Figure 2.3, clearly indicates that $\Gamma(10,5)$ has fewer long waiting times compared to $\Gamma(2,1)$.

Although patient satisfaction is an important aspect, a surgical unit is also required to treat a reasonable amount of patients (total throughput). Moreover, planned surgery should preferably be conducted within regular hours to avoid the costs of overtime. These two outcomes are considered in Papers A, D and E together with the extent of the long waiting times.

A surgical unit is highly stochastic, since the list of environmental factors influencing the system is long. This implies that also the resulting simulation model is stochastic. The model can however be put into a deterministic simulation model by keeping the seed that controls the random number generator constant. This implies that the case-study can be used for illustrating both stochastic and deterministic simulation. In the deterministic setting the model output corresponds to a single scenario and hence may not be representative for the performance in general, but the model nonetheless represents a complex deterministic simulation model.

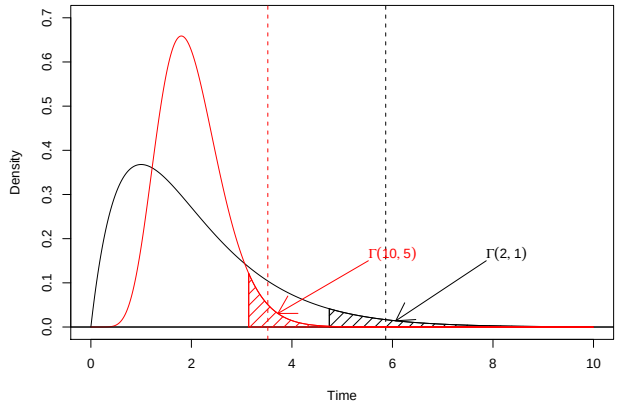


Figure 2.3: Waiting time distributions with the 5 % longest waiting times highlighted and the average waiting times of these marked by the vertical dashed lines

2.3 Queuing systems

In paper E an M/M/m-queuing system is considered, which is a system that has several appealing properties. The literature on these queuing systems is vast and their theoretical behaviour is therefore well-known and described; that is, new modeling techniques can be validated since the true input-output relation is known (as for example utilized in Kleijnen (2008) and Dehlendorff et al. (2010a)). An M/M/m-queuing system consists of a poisson arrival process and m parallel servers having exponential service times. The rate of utilization for the servers is $\rho = \lambda/(\mu m)$, where λ is the arrival rate of items (items arriving per time unit) and μ the service rate of the servers(items processed per time unit). At time points with no idle servers arriving items are queued in a queue with unlimited capacity. A typical outcome is the expected waiting time in queue, which also is the main outcome in the case-study in section 2.2 (where the queue corresponds to the delays when the resources are missing).

Figure 2.4 illustrates the outline of a M/M/4 queuing system for a hospital unit. The model in Figure 2.4 can be seen as a simplified version of the surgical unit described in section 2.2. It has four operating rooms as the model in section 2.2, but in the simplified version of the surgical unit all processes between arrival and discharge are collapsed into a queue and four parallel processes. Moreover, the M/M/4-queuing system consists of a single arrival process, whereas the surgical

unit in section 2.2, e.g., has two separate arrival processes corresponding to acute and planned patients.

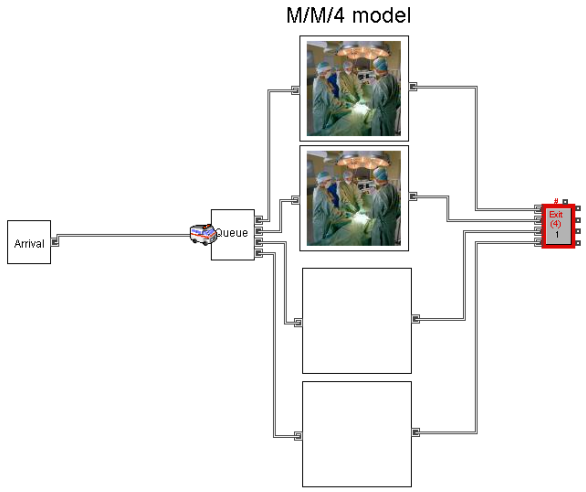


Figure 2.4: M/M/4 queue

For an M/M/m-queuing system with up to four servers the expected waiting time in the queue is given as (see e.g., Gross and Harris, 1998)

$$E[W_q] = \begin{cases} \frac{\lambda}{\mu}\frac{1}{\mu-\lambda} = \frac{\rho^2}{\lambda(1-\rho)} & m=1 \\ \frac{\lambda^2}{\mu(2\mu+\lambda)}\frac{1}{2\mu-\lambda} = \frac{2\rho^3}{\lambda(1-\rho^2)} & m=2 \\ \frac{\lambda^3}{\mu(6\mu^2+4\lambda\mu+\lambda^2)}\frac{1}{3\mu-\lambda} = \frac{9\rho^4}{\lambda(1-\rho)(2+4\rho+3\rho^2)} & m=3 \\ \frac{\lambda^4}{\mu(24\mu^3+18\lambda\mu^2+6\lambda^2\mu+\lambda^3)}\frac{1}{4\mu-\lambda} = \frac{32\rho^5}{\lambda(1-\rho)(3+9\rho+12\rho^2+8\rho^3)} & m=4 \end{cases} \quad (2.1)$$

that is; the expected waiting time in the queue can be expressed as relatively simple functions of, e.g., (λ, μ) or (λ, ρ). The relationship between ρ and W_q is visualized in Figure 2.5, which shows that with the same server utilization and arrival rate the waiting time decreases with the number of servers. This implies, e.g., that two servers with service rates μ_2 are better in terms of reducing the time spend in the queue than one twice as fast server with service rate $\mu_1 = 2\mu_2$ due to the synergy effects of two servers. For the total time spend in the system having a fast single server is better, but we only consider the waiting time in the queue.

The M/M/m-queuing system is an example of a system which can be analyzed analytically. It is however clear that if the system becomes much more complicated than this, simulation becomes the preferred method and hence conclusions

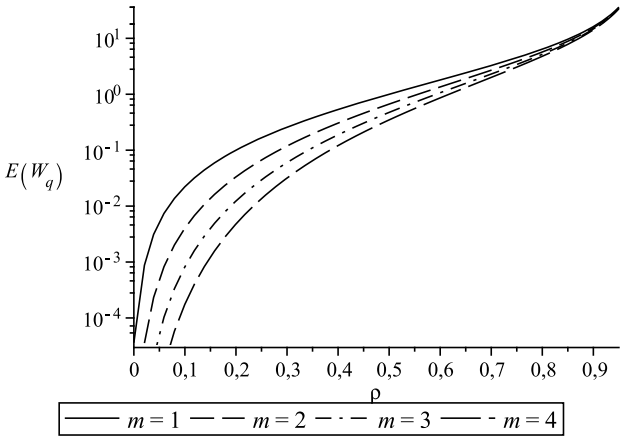


Figure 2.5: Expected waiting time in queue as function of ρ ($\lambda = 0.5$) with $m = 1, \ldots, 4$ servers

must be based on the analysis of the simulation output. This applies in many areas where the system consists of several connected components, which makes the system difficult to analyze analytically. In Paper E we use M/M/1 and M/M/2-queuing systems to illustrate three different modeling techniques for simulation models being both stochastic and influenced by uncontrollable factors.

Chapter 3

Experimental design

The relationship between input and output of a simulation or computer model is typically analyzed with a set of observations (experiments) on the model. An experimental plan (design) is a scheme for which experiments to do and in which order to run them. Such an experimental design may be organized in an $n \times s$-matrix with the ijth element containing the value of the jth of s factors in the ith of n runs. Constructing an experimental plan is a way of choosing a set of n points in the s-dimensional hypercube and many experimental design criteria are therefore based on distances between the design points in the s-dimensional design space (section 3.2 deals with optimal designs).

The first major contributions to the design and analysis of computer experiments (DACE) literature are McKay et al. (1979) and Sacks et al. (1989b), who introduce the basic foundations for DACE. In the book by Santner et al. (2003) some of the key sampling strategies and interpolation techniques are summarized. Fang et al. (2006) also discuss design and analysis of computer experiments and provide techniques for generating optimal designs. Sacks et al. (1989b) and Santner et al. (2003) consider deterministic computer experiments, i.e., computer models that generate the same output for replicated runs with the same settings of the input factors.

Experimental planning known from physical experimentation is often not well suited for deterministic computer models since, e.g., replication is deemed to

be redundant. Optimal factorial designs are popular in physical experimentation, but they are usually not applied for deterministic computer models, since projecting onto subspaces gives replicated runs; that is, if a factor turns out to be insignificant deleting this factor from the design may produce replicated runs. Consider a 2^3 full factorial design with factor B being insignificant and its projection onto factors A and C

$$\begin{bmatrix} -1 & -1 & -1 \\ +1 & -1 & -1 \\ -1 & +1 & -1 \\ +1 & +1 & -1 \\ -1 & -1 & +1 \\ +1 & -1 & +1 \\ -1 & +1 & +1 \\ +1 & +1 & +1 \end{bmatrix} \Rightarrow \begin{bmatrix} -1 & -1 \\ +1 & -1 \\ -1 & -1 \\ +1 & -1 \\ -1 & +1 \\ +1 & +1 \\ -1 & +1 \\ +1 & +1 \end{bmatrix} \tag{3.1}$$

It can be seen that the reduced design without factor B (the second column in the first design) only has four unique factor settings, which are replicated twice. Instead of using the experimental framework from physical experimentation, a separate design framework is used for computer and simulation experiments, which deals directly with the properties of these experiments.

In physical experimentation important aspects are randomization and replication (Montgomery, 2009). In computer experiments the randomization aspect is somewhat different as the random error is either not present (deterministic computer model) or controlled through a seed controlling the random number generator (stochastic computer model). Replications are for deterministic computer models redundant, since they produce the same output. Another aspect is that computer models often have many factors, complex response surfaces and long run times, which implies that typically only a very limited number of runs is affordable in a high dimensional space.

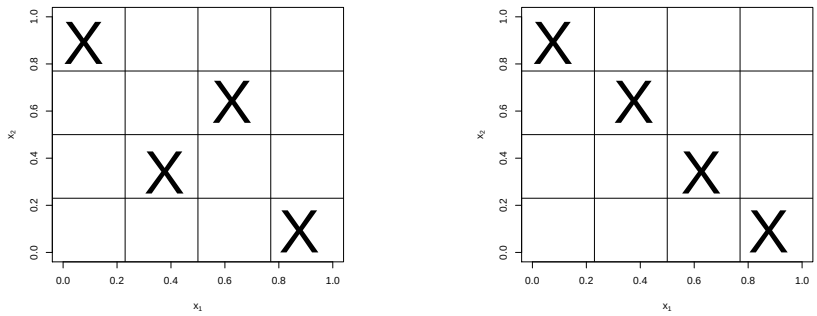
A desired property of an experimental plan for computer experiments is that the set of points chosen are space-filling (Fang et al., 2006), which implies that the design points are chosen such that they are representative for the entire design space. The space-filling requirement is motivated by the overall mean model (Fang et al., 2006), i.e., obtaining the best estimator for the overall mean of the computer model. Fang et al. (2006) state that: "... space-filling designs have a good performance not only for estimation of the overall mean, but also for finding a good approximate model". In Chapter 4 the estimation of approximate models (meta-models) is considered.

The space-filling requirement implies that the design space is required to be represented by design points in all regions and not only at, e.g., the corner points as for 2^k-factorial designs. Obviously this becomes increasingly more challenging

as the number of factors increases, i.e., the coverage of the design space tends to become sparse due to the curse of dimensionality. Another important aspect is that projecting the design onto a subset of factors should preferably result in a design without replicated runs to avoid redundant information in case of insignificant factors.

3.1 Latin hypercube sampling

A popular choice for obtaining a set of space-filling design points is latin hypercube sampling (LHS) and the associated design with n observations and s variables/factors is called a latin hypercube design (LHD(n,s)) (see for example McKay et al. (1979)). In LHS each factor's range is first divided into n intervals, which are denoted $1, \ldots, n$. For each factor a random permutation of the numbers $1, \ldots, n$ is chosen and the combination of these s permutations forms the design. For $s = 2$ and $n = 4$ one plan could be $\{3, 2, 1, 4\} \times \{3, 2, 4, 1\}$, which corresponds to the design shown in Figure 3.1(a). A different design is shown in Figure 3.1(b) and it corresponds to $\{1, 2, 3, 4\} \times \{4, 3, 2, 1\}$.



(a) (b)

Figure 3.1: LHD(3,2) experimental plans

The general constructing method for a LHD(n, s) is to combine s permutations of the numbers $1, \ldots, n$ and scale the resulting design D to the unit hypercube. The scaling can be done in multiple ways and Fang et al. (2006) consider two principal ways. The first scaling method is the midpoint latin hypercube sampling method, which for the ith run for the jth factor is given as

$$D_{ij}^{m} = \frac{D_{ij} - 0.5}{n} \tag{3.2}$$

The midpoint scaling method is used in Figure 3.1 and places the design points in the center of the squares (hypercubes in general) formed by the slicing of each factor in n intervals. The second method uses random numbers to place the design points and is given as

$$D_{ij}^{r} = \frac{D_{ij} - U_{ij}}{n} \tag{3.3}$$

where $U_{ij} \sim U(0,1)$, i.e., comes from an uniform distribution. This method places the points in each hypercube randomly instead of at its center as in midpoint scaling.

In Figure 3.1 the midpoint scaling method is used and it can be seen that projecting the design onto a single factor distributes the design points evenly with no replicates. Using the random scaling method preserves that projections do not produce replicated runs, but the distribution of design points for projections onto a single factor does not give evenly spaced points. The LHD is seen to be easy to generate, it can handle many factors and projection on to any subspace (e.g., removing a column) results in another LHD. The LHD possesses many appealing properties, however as seen from Figure 3.1 not all LHDs are equally good, e.g., the design in Figure 3.1(b) has perfectly correlated columns and hence the two factors are confounded.

3.2 Optimal designs

The problem with, e.g., correlated columns led to the development of so called optimal LHDs. Optimal LHD designs are chosen from the set of LHDs, but according to some criterion evaluating certain properties of the design. In the literature (see for example Fang et al. (2006) for a comprehensive summary) several optimality criteria are summarized, e.g., integrated mean square error (IMSE) by Sacks et al. (1989a), maximin distance by Johnson et al. (1990) and uniformity by Fang and Ma (2001). In the following it is assumed that all factors have been scaled down to $[0,1]$ and hence that the design space is the s-dimensional unit cube $[0,1]^s$.

The maximin design proposed by Johnson et al. (1990) is a design where the shortest distance between design sites is maximized

$$\max_{D} \min_{\mathbf{x_1}, \mathbf{x_2} \in D} d(\mathbf{x_1}, \mathbf{x_2}) \tag{3.4}$$

where $d()$ is a distance measure in $[0,1]^s$. The design idea is to push the design points apart such that clustering of design points is avoided, which implies that

the points are ordered such that they fill the design space. Johnson et al. (1990) also consider the minmax design

$$\min_{D} \max_{\mathbf{x}\in[0,1]^s} d(\mathbf{x}, D) \tag{3.5}$$

where $d(\mathbf{x}, D)$ is the shortest distance between $\mathbf{x}$ and the design points. The idea behind the minmax design is that any point in $[0,1]^s$ should not be too far away from a design point. The minmax design is intuitively easy to identify as being space-filling, since the criterion says that the design points should be chosen such that no region is too far away from a design point. It is however computationally much harder to find compared to the maximin design, since the maximum distance from any design point to any potential point in the design space is required.

Uniformity is another optimality criteria related to space-filling designs. It is described in great detail by Fang et al. (2006) and can be measured by, e.g., the wrap-around discrepancy (WD) as proposed by Fang and Ma (2001). The intuition behind the WD is that the fraction of design points in the hypercube spanned by any two points should match the fraction of the total volume spanned by this hypercube, which is the expected distribution of the points if they are uniformly scattered. The criteria in a computational efficient version is given as

$$(\mathrm{WD}(D))^2 = -\left(\tfrac{4}{3}\right)^s + \tfrac{1}{n}\left(\tfrac{3}{2}\right)^s + \tfrac{2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\prod_{i=1}^{s} q_i(j,k) \tag{3.6}$$

where $q_i(j,k) = \frac{3}{2} - |x_k^i - x_j^i|(1 - |x_k^i - x_j^i|)$, n is the number of points, s is the number of factors (the dimension), and x_k^i is the ith coordinate of the kth point. A low WD value corresponds to a high degree of uniformity. Since $x_k^i \in [0,1]$, $q_i(j,k)$ is maximal when the distance between x_k^i and x_j^i is either 0 or 1 and minimal with a distance of 0.5. The wrap around part of the criteria arises since the hypercube spanned by two design points may potentially wrap around the bounds of the unit cube, which is illustrated by the highlighted area in Figure 3.2. The L_2 relates to how the discrepancy between the fraction of points contained in the hypercube spanned by two design points and its volume is measured. L_2 is simply the squared difference, which is given as

$$\left|\frac{\text{number of points in hypercube}}{\text{total number of points}} - \text{Volume of hypercube}\right|^2 \tag{3.7}$$

Other measures exist, such as the centered discrepancy, which however depends on the corner points, whereas the wrap-around discrepancy is said to be unanchored. Fang et al. (2006) points out that there is a connection between orthogonal designs and uniform designs for example that "any orthogonal design is a uniform design under a certain discrepancy".

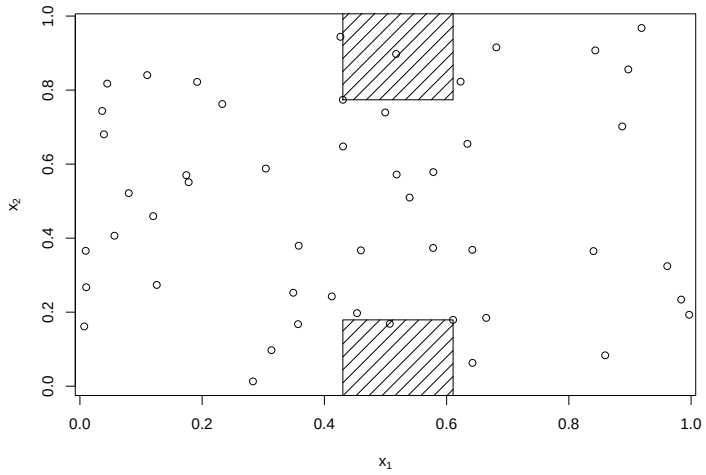


Figure 3.2: Illustration of wrap-around discrepancy

In Papers B and C uniform designs are used, since they according to Fang et al. (2006) are robust against the a priori model assumption for the meta-model, i.e., they do not rely on a specific model structure. The uniform designs can be generated by the good lattice point method described in Fang et al. (2006). The construction of the design is based on a lattice $\{1, \ldots, n\}$ and a generator $\mathbf{h}(k) = (1, k, k^2, \ldots, k^{s-1})(\text{mod } n)$, with k fulfilling that $k, k^2, \ldots, k^{s-1}(\text{mod } n)$ are distinct. The generator $\mathbf{h}(k)$ is chosen such that the resulting design consisting of the elements $u_{ij} = ih(k)_j(\text{mod } n)$ scaled down to $[0, 1]^s$ has the lowest WD value.

3.3 Crossed designs

In some simulation applications the input factors of the model consist of both controllable and uncontrollable factors. This implies that a different experimental design strategy is needed, since the two factor types have different roles and interpretation in the physical system. For example optimization of the performance of the system only involves choosing the best combinations of the controllable factors, since in the physical system the uncontrollable factors can not be fixed at certain values. However, the performance of the system may depend on the settings of the uncontrollable factors, which implies that several

settings of the uncontrollable factors must be tested at each setting of the controllable factors in order to ensure that conclusions based on the controllable factors are robust.

Crossed designs are used for combining two or more designs. In particular in applications with controllable and uncontrollable factors this method is used to test the controllable factor settings under different uncontrollable factor settings (Kleijnen, 2008, 2009). One could for example consider a factorial design for the controllable factors and a LHD for the uncontrollable factors and obtain a combined design by crossing the two designs. This is illustrated by the following example

$$\begin{bmatrix} -1 & -1 \\ +1 & +1 \end{bmatrix} \times \begin{bmatrix} 1 & 2 & 4 \\ 2 & 3 & 2 \\ 3 & 1 & 1 \\ 4 & 4 & 3 \end{bmatrix} \Rightarrow \begin{bmatrix} -1 & -1 & 1 & 2 & 4 \\ -1 & -1 & 2 & 3 & 2 \\ -1 & -1 & 3 & 1 & 1 \\ -1 & -1 & 4 & 4 & 3 \\ +1 & +1 & 1 & 2 & 4 \\ +1 & +1 & 2 & 3 & 2 \\ +1 & +1 & 3 & 1 & 1 \\ +1 & +1 & 4 & 4 & 3 \end{bmatrix} \tag{3.8}$$

which shows the result of crossing a 2^{2-1} fractional factorial design with a LHD(4,3) (the low and high levels of the factors in the factorial design are coded "-1" and "$+1$", respectively).

It can be argued that crossing two designs may not be the optimal way of choosing the settings for the uncontrollable factors, since the settings of the uncontrollable factors are replicated n_c times each. Covering the uncontrollable factor space is important in order to obtain a better understanding of the uncontrollable factors and to ensure that important uncontrollable factor effects are not overlooked. Moreover, since the specific setting of the uncontrollable factor is not of interest, then more information from the simulation model is obtained by using different settings of the uncontrollable factors for each setting of the controllable factors. One challenge is to construct the sub-designs such that they are similar, i.e., that the controllable factor settings are exposed to the same range of uncontrollable factor settings. This is achieved by the design we propose in section 3.4.

3.4 Top-Down design

The replications of the uncontrollable factor settings in the crossed design inspired us to develop a different experimental plan, which is presented in Papers B

Controllable factor setting	*Top-down design* Uncontrollable factor setting				*Crossed design* Uncontrollable factor setting			
x_{c_1}	x_{e_1}	x_{e_2}	x_{e_3}	x_{e_4}	x_{e_1}	x_{e_2}	x_{e_3}	x_{e_4}
x_{c_2}	x_{e_5}	x_{e_6}	x_{e_7}	x_{e_8}	x_{e_1}	x_{e_2}	x_{e_3}	x_{e_4}
x_{c_3}	x_{e_9}	$x_{e_{10}}$	$x_{e_{11}}$	$x_{e_{12}}$	x_{e_1}	x_{e_2}	x_{e_3}	x_{e_4}
x_{c_4}	$x_{e_{13}}$	$x_{e_{14}}$	$x_{e_{15}}$	$x_{e_{16}}$	x_{e_1}	x_{e_2}	x_{e_3}	x_{e_4}
x_{c_5}	$x_{e_{17}}$	$x_{e_{18}}$	$x_{e_{19}}$	$x_{e_{20}}$	x_{e_1}	x_{e_2}	x_{e_3}	x_{e_4}

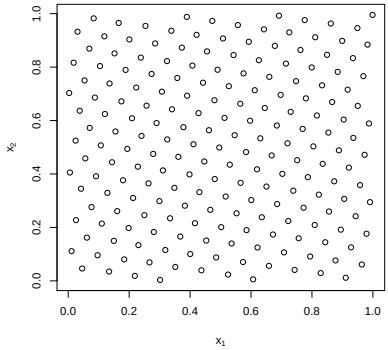
Table 3.1: Top-down design with $n_c = 5$ and $n_u = 4$ compared to a crossed design of same size

and C. In this design different uncontrollable factor settings are used for each controllable factor setting and has a "top-down" structure and hence denoted a top-down design (Dehlendorff et al., 2008, 2011).

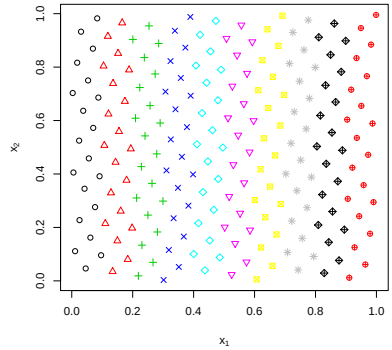
The construction of the top-down design is illustrated in Figure 3.3 and it consists of five steps:

1. construct a uniform design for the uncontrollable factors with $n = n_c \times n_u$ runs (Figure 3.3(a)), where n_c is the size of the design for the controllable factors and n_u is the number of uncontrollable factor settings to test at each setting of the controllable factors.
2. split the overall design into n_u initial subregions (Figure 3.3(b))
3. add n_u center points (Figure 3.3(c))
4. permute the assignment of points such that the subregions are well defined/more compact (Figure 3.3(d))
5. assign each controllable factor setting one point from each subregion such that all points are assigned to a controllable factor setting (Figure 3.3(e)).

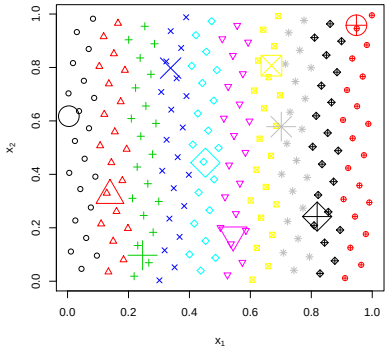
The benefit of using the top-down design compared to the crossed design is that n_c as many different settings of the uncontrollable factors are tested, which implies that the uncontrollable factor space has a higher coverage. The higher coverage is in Paper C shown to reveal important interactions between controllable and uncontrollable factors, which may be used to put the system in a more robust operating mode. The main challenge in the construction method is to assign the uncontrollable factor settings such that the variations in the uncontrollable factors (corresponding to the environment) is comparable from one setting of the controllable factors to the next. The top-down design is described in greater detail in the summaries of Papers B and C in sections 5.2 and 5.3.



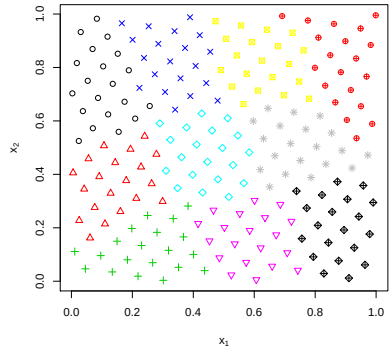
(a) First construct an uniform design ($n = n_c \times n_u$



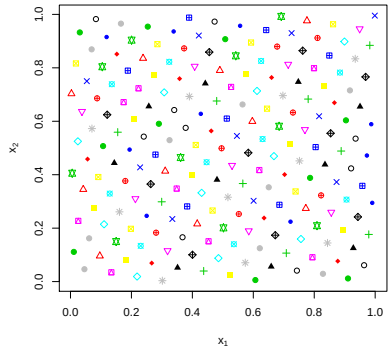
(b) Divide the design into n_u sub-regions consisting of n_c points



(c) Add n_u center points



(d) Reorganize points into n_u well defined sub-regions around the center points



(e) Assign one point from each subregion to each controllable factor setting

Figure 3.3: Top-down algorithm

Chapter 4

Output analysis

An often occurring challenge with computer and simulation models is that they can be very expensive in terms of the time it takes to complete a single run. This implies that the models are not well suited for optimization, since this usually requires many evaluations. For computational expensive computer models an often used technique is therefore to build a computationally cheaper model called a meta-model. A meta-model is thus an approximation of the input-output relationship of the computer model (Santner et al., 2003, Fang et al., 2006, Kleijnen, 2009).

In this thesis two groups of analysis methods are considered: Kriging and regression models. Kriging (Matheron, 1963) is the preferred model for deterministic simulation and computer models, since it interpolates the observations (see section 4.1). Regression models as described in section 4.2 are extensively used in the analysis of physical experiments, but can also be used for stochastic simulation and computer models. In section 4.3 we give a small example of how a computer model can be optimized using a meta-model.

4.1 Kriging

A natural requirement for meta-models for deterministic computer models is that they interpolate the data, i.e., that the meta-model equals the computer model at the design sites. A popular modeling framework is Kriging, which originates from geo-statistics. The method was developed by Krige and improved by Matheron (1963) and is often applied in the field of computer experiments (Sacks et al., 1989b, Santner et al., 2003, Martin and Simpson, 2005, Kleijnen, 2009). The method has several advantages 1) the predictor interpolates the data points, 2) the model is global and 3) it can fit complex response surfaces. However using the model outside the data range is known to give poor predictions as noted by van Beers and Kleijnen (2004).

We consider a function or model that, given the input vector $\mathbf{x}$, generates the scalar and deterministic output $y(\mathbf{x})$. The Kriging model relies on the assumption that the deterministic output $y(\mathbf{x})$ can be described by the random function

$$Y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \boldsymbol{\beta} + Z(\mathbf{x}) \tag{4.1}$$

where $\mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}$ is a parametric trend with p parameters and $Z(\mathbf{x})$ is a random field assumed to be second order stationary with covariance function $\sigma^2 R(\mathbf{x_i}, \mathbf{x_j})$ (Santner et al., 2003), where σ^2 is the variance and $R()$ is the correlation function, which usually is assumed to be the gaussian correlation function given as

$$R(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(- \sum_{j=1}^{p} \theta_j (x_1^j - x_2^j)^2 \right) \tag{4.2}$$

where x_i^j is the value of the jth factor of observation i and $\theta_j \geq 0$ the corresponding correlation parameter. $\theta_j = 0$ implies that the correlation along the jth factor is 1.

We consider a set of n design points $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and corresponding observations $\mathbf{y} = \{y(\mathbf{x_1}), \ldots, y(\mathbf{x_n})\}$ where $y()$ is the true function (computer model). The correlation matrix for the design points is denoted $\mathbf{R}(\boldsymbol{\theta})$ where the ijth element is the correlation between the ith and jth design points given as $R(\mathbf{x_i}, \mathbf{x_j})$. Likewise the vector of correlations between the point, $\mathbf{x}$, and the design points is defined as

$$\mathbf{r}(\mathbf{x}) = [R(\mathbf{x_1}, \mathbf{x}), \ldots, R(\mathbf{x_n}, \mathbf{x})]^T \tag{4.3}$$

The regressor $\mathbf{f}(\mathbf{x})$ is given by a vector with p regressor functions

$$\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}) \ldots f_p(\mathbf{x})]^T \tag{4.4}$$

and the regressors for the design sites are given as

$$\mathbf{F} = [\mathbf{f}(\mathbf{x_1})^T \cdots \mathbf{f}(\mathbf{x_n})^T]^T \tag{4.5}$$

Usually ordinary Kriging is used and hence $\mathbf{f}(\mathbf{x})$ reduces to $f(x) = 1$ corresponding to the model

$$Y(\mathbf{x}) = \mu + Z(\mathbf{x}) \tag{4.6}$$

The correlation function is parameterized by a set of parameters $\boldsymbol{\theta}$ as described in (4.2). Given $\boldsymbol{\theta}$, the restricted maximum likelihood estimate of $\boldsymbol{\beta}$ (Santner et al., 2003) (assuming a gaussian distribution) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^T \hat{\mathbf{R}}(\boldsymbol{\theta})^{-1} \mathbf{F})^{-1} \mathbf{F}^T \hat{\mathbf{R}}(\boldsymbol{\theta})^{-1} \mathbf{y} \tag{4.7}$$

where $\hat{\mathbf{R}}(\boldsymbol{\theta})$ is the correlation matrix for the design sites and parameterized by the parameter vector $\boldsymbol{\theta}$. The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})^T \hat{\mathbf{R}}(\boldsymbol{\theta})^{-1} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) \tag{4.8}$$

where n is the number of observations and p is the rank of F (the number of parameters in $\hat{\boldsymbol{\beta}}$). The correlation parameters are found by minimizing the negative restricted profile log-likelihood (L_r) for $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[(n-p) \log \hat{\sigma}^2 + \log(|\mathbf{R}(\boldsymbol{\theta})|) \right] \tag{4.9}$$

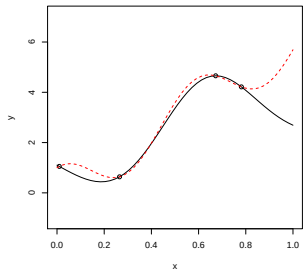
where $|\mathbf{R}(\boldsymbol{\theta})|$ is the determinant of the correlation matrix corresponding to the design points. $\hat{\sigma}$ and $\hat{\boldsymbol{\beta}}$ are functions of $\hat{\mathbf{R}}^{-1}$ (equation (4.7) and (4.8)); that is, inverting the correlation matrix for the design sites is required in order to evaluate the likelihood function. This inversion is a computational expensive task since it takes $O(n^3)$ operations. Moreover, the likelihood function may be flat around the optimum, which implies that the search for the optimum may become slow (Lophaven et al., 2002a, Li and Sudjianto, 2005). These aspects are dealt with in the Matlab toolbox `DACE` by Lophaven et al. (2002b).

Given $\hat{\mathbf{R}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ the predictor at $\mathbf{x}$ is

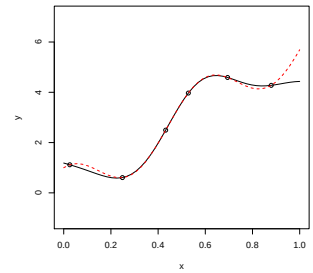
$$\hat{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x})^T \hat{\mathbf{R}}^{-1} (\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) \tag{4.10}$$

At a design point, $\mathbf{x} \in \mathbf{X}$, the vector $\mathbf{r}(\mathbf{x})^T \hat{\mathbf{R}}^{-1}$ consists of $(n-1)$ zeroes and a single one at the index corresponding to $\mathbf{x}$, which implies that the predictor becomes $y(\mathbf{x})$ and thus interpolates the data at the design points. The interpolation property is one of the main advantages of using Kriging for deterministic computer models.

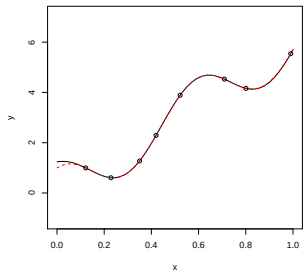
An example of the Kriging predictor is shown in Figure 4.1. It can be seen that the interpolator is improving as more design points are added, i.e., the difference between the interpolator and the true function is not visible for $n = 10$ design points (Figure 4.1(d)). The performance of the predictor can be measured by the accuracy, $1/(1 + \text{RMSE})$, where RMSE is the root mean square prediction error over a set of test sites. The accuracy is in Figure 4.1 seen to increase as the number of design points is increasing. Likewise the correlation between points is seen to increase ($\hat{\theta}$ is decreasing) as more design points are included. It can be seen that the interpolator is able to fit a quite wiggly curve using only two parameters: $\hat{\beta}$ and $\hat{\theta}$.



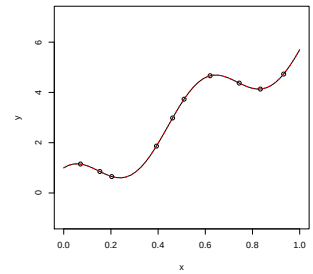
(a) Kriging interpolator based on 4 design points ($1/(1 + \text{RMSE}) = 0.56$, $\hat{\theta} = 1.60$)



(b) Kriging interpolator based on 6 design points ($1/(1 + \text{RMSE}) = 0.79$, $\hat{\theta} = 1.01$)



(c) Kriging interpolator based on 8 design points ($1/(1 + \text{RMSE}) = 0.96$, $\hat{\theta} = 0.60$)



(d) Kriging interpolator based on 10 design points ($1/(1 + \text{RMSE}) = 1.00$, $\hat{\theta} = 0.50$)

Figure 4.1: Illustration of Kriging predictor for 4-10 points. Solid black lines correspond to the true function, dashed red lines are the Kriging predictors and "o" corresponds to the design points. The underlying signal is $y = \cos(6.8\pi x/2) + 6x$

4.2 Regression models

If the output of the computer model is stochastic, an interpolator such as the Kriging model may not be the best predictor (see for example Figure 1.3(b)). Instead regression methods from physical experimentation can be applied. However, one difference is that in simulation the random error is usually controlled through the seed to the random number generator, which implies that the observations may not be independent. In such cases, e.g., generalized least squares methods can be used (Kleijnen, 2008). In this thesis we however only consider experiments with the seed either kept fixed (deterministic simulation) or chosen randomly for each run (stochastic simulation).

In the following we consider the most general simulation model, which is stochastic and has controllable and uncontrollable factors. Let x_i^c be the ith controllable factor setting, x_j^u the jth uncontrollable factor setting and s_{ijk} the seed in the ijkth run. Moreover, we focus on modeling the variation coming from the uncontrollable factors and the seed, i.e., consider the combinations of the settings of the controllable factors as a single categorical variable to simplify the analysis and focus on the uncontrollable factors.

A simple model for stochastic simulation is the general linear model, i.e., the model

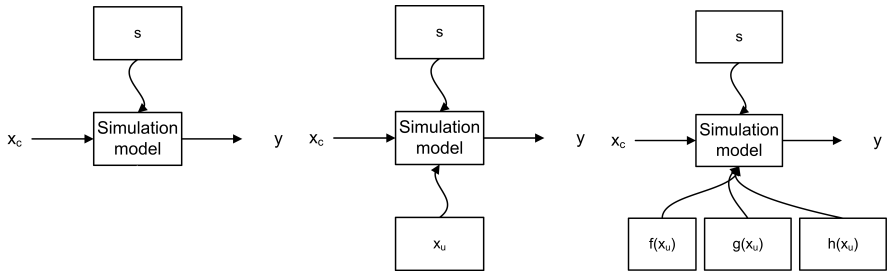
$$y(x_i^c, x_j^u, s_{ijk}) = \beta_i + \epsilon_{ijk} \tag{4.11}$$

where β_i is the parameter for the ith controllable factor setting and $\epsilon_{ijk} \sim N(0, \sigma^2)$. In equation (4.11) the variation due to the uncontrollable factors is ignored and pooled into a single variance component together with the variation due to the seed. The variation coming from changes in the uncontrollable factors can be estimated by fitting a linear mixed effects model, which is given as

$$y(x_i^c, x_j^u, s_{ijk}) = \beta_i + U_j + S_{ijk} \tag{4.12}$$

In the linear mixed effects model the variation due to the uncontrollable factors is captured in $U_j \sim N(0, \sigma_U^2)$, whereas the variation due to the seed is captured in $S_{ijk} \sim N(0, \sigma_S^2)$. U_j and S_{ijk} are assumed to be independent, which implies that the variance of a single test/run can be written as $\sigma^2 = \sigma_U^2 + \sigma_S^2$.

In Paper C a generalized additive model (Hastie and Tibshirani, 1990, Wood, 2006) is applied to the output from a top-down and a crossed experiment on the simulation model for the surgical unit. The model is also used in Papers D and E as an extension to the linear and linear mixed effects models. The generalized additive model (GAM) is given as a function of both controllable and



(a) Linear model (b) Linear mixed effects model (c) GAM model

Figure 4.2: Illustration of models for output from stochastic simulation model with controllable and uncontrollable factors

uncontrollable factors

$$y(x_i^c, x_j^u, s_k) = \beta_i + \sum_{l=1}^{m} f_l(x_j^{u(l)}) + S_{ijk} \tag{4.13}$$

with $x_j^{u(l)}$ being the jth setting for the lth uncontrollable factor and $S_{ijk} \sim N(0, \sigma_S^2)$ the residual or seed term. f_l is a spline based smooth function with the smoothness determined by a penalty term. By estimating the functional relationship between the uncontrollable factors and the outcome, the uncontrollable factors that are needed to be tightly controlled may be identified. But more importantly interactions between controllable and uncontrollable factors may also be estimated by fitting different smooth functions depending on the settings of the controllable factors. The interactions between controllable and uncontrollable factors may be used to put the system in a more robust operating mode as suggested by Bursztyn and Steinberg (2006) and Myers et al. (2009). The estimation of the β's and the smooth functions can for example be done with the `R`-library (R Development Core Team, 2007) provided by Wood (2006).

A graphical overview of the three models is given in Figure 4.2, which shows that the models have increasingly more structure for the uncontrollable factors. The models may also be expanded by putting more structure in the controllable factor part, e.g., including low order polynomials to account for the effects of the controllable factors. In this thesis we, however, primarily focus on describing the variations in the uncontrollable factors. For all three models generalized versions exist such that, e.g., binomial and count data can be fitted. The generalized versions are considered in Paper D for estimating the risk of putting the surgical unit in a worse operating mode compared to the current setting.

4.3 Example: Optimization using a meta-model

If the computer or simulation model is too expensive to use directly for optimization a meta-model can be used as a replacement of the expensive model. Optimization can, e.g., be done in the following four stages

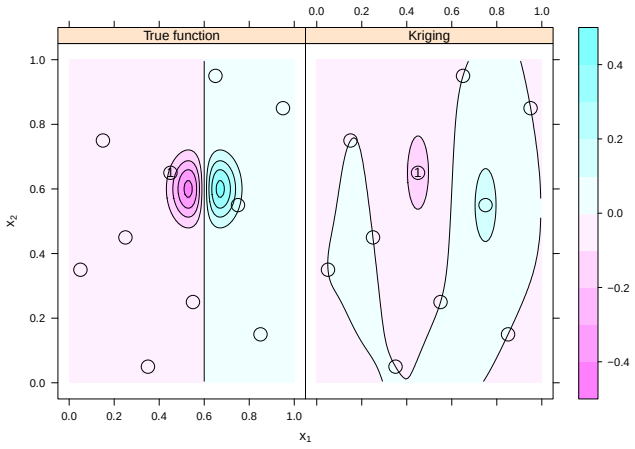
1. run initial design on expensive computer model
2. fit a meta-model based on the observations from the initial design
3. optimize the system using the meta-model
4. validate the optimal setting by running a small number of control runs on the computer model (and possibly return to the second step after adding more observations if optimum is not reached)

Using the meta-model not only speeds up the optimization but may also increase the understanding of the complex computer model if the simpler meta-model has a more explicit relationship between the input factors and the output (provided that the meta-model is an adequate description). However, using a meta-model assumes that the optimum is within the design region (local optimization), whereas the response surface methodology is generally preferred for global optimization (see for example Myers et al., 2009).

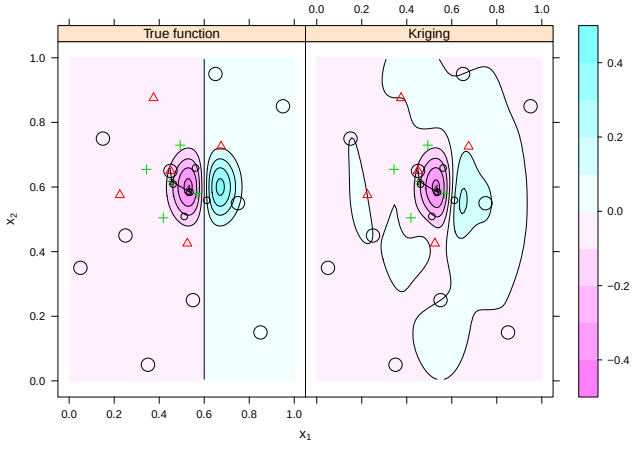
We now illustrate optimization using a meta-model by a small example with a known function, which is given as $y(x_1, x_2) = (10x_1 - 6)\exp[-(10x_1 - 6)^2 - (10x_2 - 6)^2]$ for $(x_1, x_2) \in [0, 1]^2$. A contour plot of the true function is shown in Figure 4.3, which shows that the function is mostly flat and has its maximum and minimum in the same proximity. The objective of the optimization is to find the minimum of the function $y(x^*) = y(x_1^*, x_2^*)$ by using a meta-model for the optimization task. In this example a Kriging model is used, since the output is deterministic.

First an initial maxmin LHD(10,2) is constructed and then the computer model run for these ten settings. This gives a set of observations $y^1, \ldots, y^{10}$ at the design sites $(x_1^1, x_2^1), \ldots, (x_1^{10}, x_2^{10})$ for which a Kriging model is fitted. Optimization can then be done by evaluating the Kriging predictor over a fine grid of say 10.000 points or by using standard optimization software, e.g., `optim` in *R* (R Development Core Team, 2007). This gives the estimated minimum $\hat{x}^*$ with the predicted value $\hat{y}(\hat{x}^*)$.

The estimated minimum, $\hat{x}^*$, based on the initial ten points is marked by "1" in Figure 4.3(a). It can be seen that $\hat{x}^*$ is in the neighborhood of the true minimum,



(a) 10 initial data points (maxmin LHD)

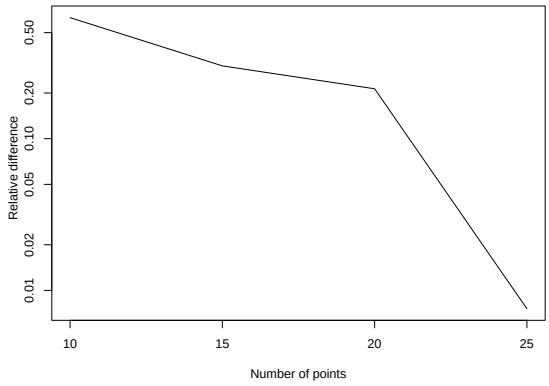


(b) After 15 additional data points

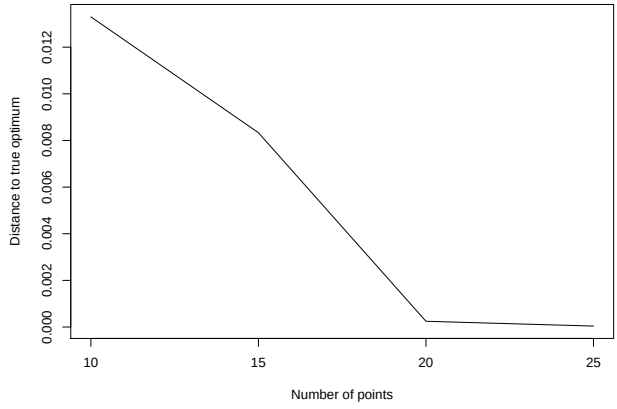
Figure 4.3: Optimizing computer model by using a meta-model. a) shows the initial model to the right and the true function to the left. The estimated optimum is marked with "1" and the data points with "O". b) shows the model after three iterations with the estimated optimums marked by connected lines.

but still not entirely correct. The relative difference between $y(\hat{x}^*)$ and $\hat{y}(\hat{x}^*)$ (the difference between the true function value at the estimated minimum and the estimated function value at the estimated minimum) is more than 50 % (Figure 4.4(a)).

To improve the estimated minimum new points are added and evaluated by the true function and the Kriging model and x^* are updated until the relative difference between $y(\hat{x}^*)$ and $\hat{y}(\hat{x}^*)$ is under 1 %. In this example we add four new points around x^* and reuse the already calculated value at the estimated minimum (calculated for the evaluation of the estimated minimum). It can be seen from Figure 4.4 that after 15 additional points the difference between the estimated and true minimum is small in both location and function value. Actually the estimated optimums are close in location after 10 additional points, but the predicted value is not. If the computer code is very time consuming, this method may give huge savings in computing time, since the Kriging model is very cheap to evaluate. This is also utilized by Dellino et al. (2009) to find robust solutions in simulation by using methods inspired by Taguchi (Taguchi, 1987).



(a) Relatively difference between $y(\hat{x}^*)$ and $\hat{y}(\hat{x}^*)$



(b) Distance to true minimum

Figure 4.4: Improvement in Kriging estimator for the minimum of the function considered in Figure 4.3 in terms of function value 4.4(a) and location 4.4(b)

Chapter 5

Summary of papers

5.1 Paper A

Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units

The topic of Paper A is comparison of statistics describing waiting time distributions. In health care applications patient waiting time is a frequently occurring measure of quality. The objective is therefore to summarize a sample of waiting times, $T = t_1, \ldots, t_N$, such that certain properties are highlighted. The background of the paper is the simulation model in section 2.2 for which reducing long waiting times for the patients is an important performance parameter. Avoiding or reducing long waiting times is important since according to Bielen and Demoulin (2007) patient satisfaction decreases as the waiting time increases.

Several statistics for samples of waiting times such as the average and maximum waiting time are used in the literature. In Paper A we propose Conditional Value of Risk (CVaR) (Kibzun and Kuznetsov, 2003, 2006) as a measure of the extent of long waiting times. CVaR originates from economics where it is used in, e.g., portfolio management as a measure of risk. For waiting times it becomes a measure of the risk of long waiting times, which is an important parameter in terms of patient satisfaction (Bielen and Demoulin, 2007). Often waiting time

distributions are right skewed consisting of mainly short waiting times, but may also have long tails corresponding to the less frequently occurring long waiting times.

The average waiting time taken over all patients corresponds to disregard the distribution of the waiting times and only focus on the overall waiting time. This is in economics known to be a risk neutral strategy, i.e., it only considers the expected loss and not the risk of big losses. Another measure is the maximum waiting time, which is seen to belong to the other extreme where the shape of the distribution once again is ignored but now only the longest waiting time is used. Using the maximum is in economics known as a risk averse strategy. The maximum waiting time is also a problematic statistic, since it is a measure of an extreme (it relies on a single observation); that is, the uncertainty of the maximum waiting time is high and hence may require a large sample and many replications to estimate properly. Moreover, it may be a too restrictive strategy and may also not represent the performance of the system, e.g., be an extremely rare observation in an otherwise well performing system.

In Paper A we propose CVaR as a compromise between these two extremes. CVaR is the average of the $(1-\alpha)100\%$ longest waiting times and is given as

$$CVaR_\alpha(T) = \frac{1}{1-\alpha}\left[\left(\frac{i_\alpha}{N}-\alpha\right)t_{i_\alpha} + \sum_{i=i_\alpha+1}^{N}\frac{t_i}{N}\right] \tag{5.1}$$

where α is the level of risk aversion, $t_1 \leq t_2 \leq \cdots \leq t_N$ are the ordered waiting times, i_α is the index satisfying $\frac{i_\alpha}{N} \geq \alpha > \frac{i_\alpha-1}{N}$ (the α-percentile) and N is the sample size. It can be seen that $CVaR_0(T) = \bar{T}$ (the average waiting time) and $\lim_{\alpha\to 1} CVaR_\alpha(T) = \max_{i=1,\ldots,N} t_i$ (the maximum waiting time). CVaR can therefore be seen as a compromise between the average and the maximum waiting time and α determines the relative importance of the longest waiting times or the level of risk aversion. A related measure is the Value at Risk waiting time (VaR), which is given as $VaR = t_{i_\alpha}$. It is however generally not recommended, since it is not sensitive to the shape of the distribution of the $(1-\alpha)100\%$ longest waiting times.

The benefits of using CVaR are illustrated by a simulation model of an orthopedic surgical unit. The model was developed in collaboration with Gentofte University Hospital, Copenhagen. The paper consists of two examples; in the first example the porter resource is varied from one to four porters and in the second example the volume of the elective patients is increased by 7, 14 and 29 % while the number of porters is kept constant at four. The examples illustrate that the average waiting time is not always the best statistic since it may overlook important shifts in the tail of the waiting time distribution. Figure 5.1 and 5.2 show that the absolute changes in CVaR are larger compared to the

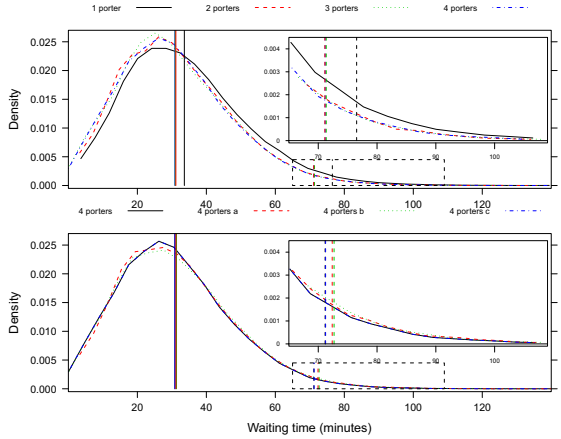


Figure 5.1: Estimated densities for seven different scenarios: 1-4 porters (top) and 4 porters with 7, 14 and 29 % more elective patients (bottom). The average waiting times are marked with solid vertical lines, whereas the CVaR waiting times are marked with dashed vertical lines.

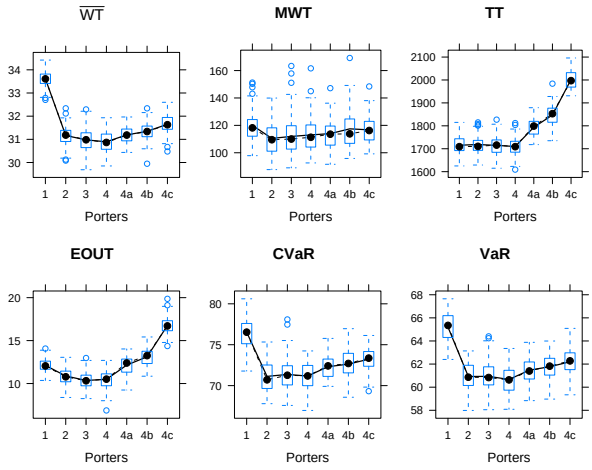


Figure 5.2: Comparison of six different performance measures for seven different scenarios: 1-4 porters and 4 porters with 7 % (4a), 14 % (4b) and 29 % (4c) more elective patients. $\bar{WT}$ is the average waiting time, MWT is the maximum waiting time, TT is the total throughput, $EOUT$ the percentage of elective patients treated outside regular hours, $CVaR$ is the CVaR waiting time and VaR is the VaR waiting time.

average waiting time, since CVaR is more sensitive to changes in the tail of the waiting time distribution.

Figure 5.2 furthermore shows that using the maximum waiting time may be problematic due to the uncertainty of this statistic; that is, the maximum waiting time is close to being the same regardless the number of porters and elective patient volume. The example shows that the compromise between the average waiting time and the maximum waiting time given by the CVaR waiting time is a reliable measure for measuring the extent of long waiting time.

Dellino et al. (2009) use constrained optimization, i.e., they optimize the mean given a standard deviation constraint. This leads to the so-called Pareto-optimal frontier, i.e., a curve showing the relationship between the risk (standard deviation) and the profit (the mean). They fit separate Kriging models for the mean and for the standard deviation and use bootstrapping to estimate regions of confidence for the mean and standard deviation given a specific constraint. As also mentioned by the authors, CVaR may be used as replacement of the mean-variance technique.

5.2 Paper B

Designing Simulation Experiments with Controllable and Uncontrollable Factors

In Paper B design of simulation experiments with two types of factors (controllable and uncontrollable) is considered. The two factor types have different interpretation in the physical system and hence need to be treated differently; that is, the system is optimized in the controllable factors such that the setting is optimal disregarding the settings of the uncontrollable factors. The experimental design is therefore required to be run under various settings of the uncontrollable factors for each combination of the controllable factors.

Models with controllable and uncontrollable factors are often analyzed using a crossed design (Kleijnen, 2008). This implies that the same combinations of settings for the uncontrollable factors are used for all combinations of the controllable factor settings (whole plots) and hence that the uncontrollable factor space is sparsely covered due to the replications as discussed in section 3.3. It could therefore be argued that using different settings of the uncontrollable factors for each whole plot is a better way of choosing the settings of the uncontrollable factors. For n_c whole plots this gives n_c as many different uncontrollable factors combinations, i.e., a higher coverage of the uncontrollable factor space.

The main challenge in designing such an experimental plan is to make the subdesigns for the uncontrollable factors similar from one whole plot to the next while ensuring that the overall design is uniform. In Paper B this is achieved in two different ways. The first strategy has a bottom-up structure and the design is constructed from n_u regions each consisting of n_c space-filling points (see Figure 5.3).

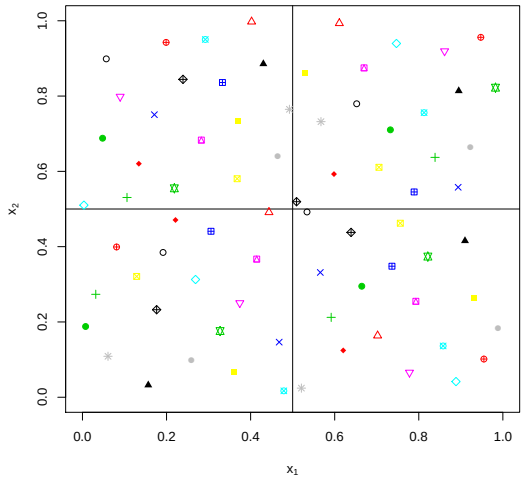


Figure 5.3: Illustration of bottom-up design with four subregions

The whole plots are then assigned one design point from each of the n_u regions such that all points are assigned. However, the bottom-up strategy does not guarantee the uniformity of the combined design, which can be seen from Figure 5.4. The best bottom-up design with 200 runs (five controllable factor settings each with 40 uncontrollable factor settings) for two uncontrollable factors is seen to have a WD-value approximately five times higher than an uniform design generated directly.

Instead we propose a second strategy, which has more of a top-down structure where the overall design is constructed first to guarantee the overall uniformity (see section 3.4). The overall design is then split into subdesigns one for each whole plot. The subdesigns are generated by splitting the $N = n_u n_c$ points into n_u subgroups of n_c points and then assigning each whole plot one point from each subgroup. The assignment of points can be done in many ways and the WD-values of the subdesigns are used as criteria for the best assignment, we choose the assignment where the maximum WD-value of the subdesigns

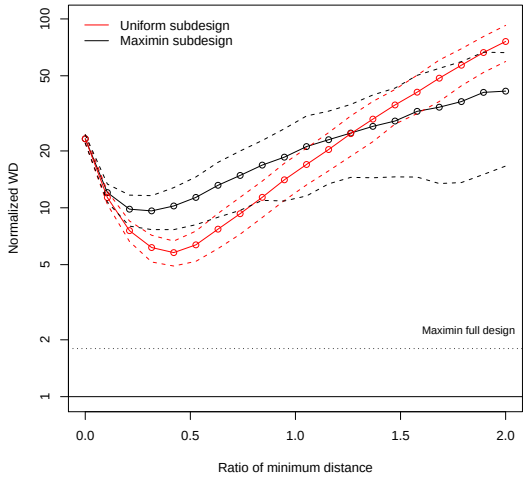


Figure 5.4: Uniformity of combined design with bottom-up strategy

is lowest. In Paper C the top-down design is considered in more detail and compared to the crossed design using the simulation model from section 2.2.

The main contribution in Paper B is the development of an experimental plan giving a high coverage in the uncontrollable factor space for simulation models having both controllable and uncontrollable factors. In paper C we show that the higher coverage leads to a better understanding of the uncontrollable factors.

5.3 Paper C

Designing simulation experiments with controllable and uncontrollable factors for applications in health care

In Paper C we reconsider the proposed experimental design in Paper B. The benefit of using the top-down design is illustrated by the simulation model described in section 2.2 (see also Paper 5.1). The top-down design is compared with the crossed design (see equation 3.8), which is the most commonly used design for simulation experiments with controllable and uncontrollable factors. The output is analyzed with generalized additive models (Hastie and Tibshirani, 1990, Wood, 2006) for both of the considered experiments (see section 4.2).

The model output is modeled by the GAM model, i.e., a flexible regression method. In the paper it is shown that the top-down design identifies important interactions between the controllable and uncontrollable factors, which in the example is not identified using the crossed design (see Figure 5.5). These interactions are important, since they may be used to put the system in a robust operating mode.

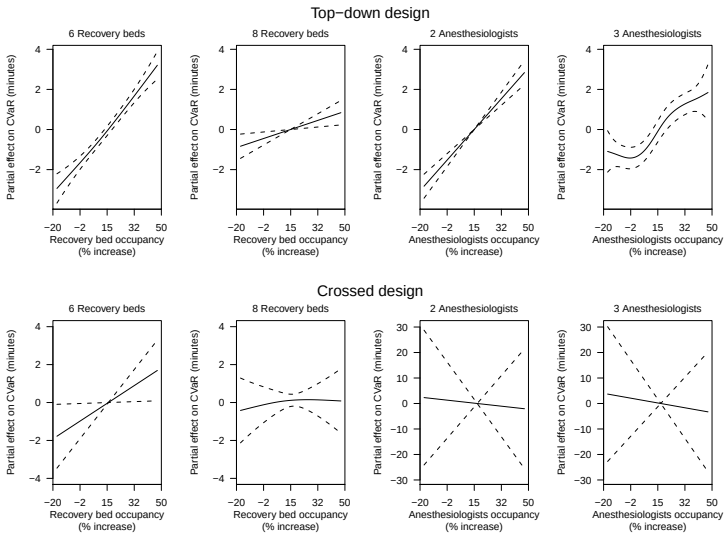


Figure 5.5: Interactions between controllable and uncontrollable factors

The top-down design may also be used as a method for generating a sequential sampling scheme in the following manner: disregard the controllable/uncontrollable setup, instead we consider the top-down design as n_c batches of runs, which are run sequentially one batch at the time. This may give a faster completion of the experiment if not all batches are needed. However, this only works in the simple case with only one type of factors in which the controllable factors settings correspond to batches and the uncontrollable factors to the factors of the model. Kleijnen and van Beers (2004) also consider sequential sampling using Kriging as a meta-model, which is extended in van Beers and Kleijnen (2008) who consider sequential sampling for random simulation. Sequential sampling fits very well with simulation, since the simulation experiments are run sequentially. Strategies for generating the next sampling point and/or stopping the procedure can therefore be implemented between two runs or between batches of runs.

A modification of the top-down design is to consider a different distribution of

the points, i.e., instead of an uniform distribution in each dimension, it may be more relevant to spread the points out corresponding to a gaussian distribution. In such a design the emphasis is put on the center of the gaussian distribution corresponding to that certain regions are of greater importance than others, e.g., a-priori knowledge lead us to believe that the optimum or the function is highly variable in these regions. The uniform design spread the points evenly on each factor, which can be transformed to a gaussian distribution in the following way

1. Construct a top-down design with $N = n_c n_u$ runs and p uncontrollable factors and denote the settings of i'th uncontrollable factor $x^i = [x_1^i, \ldots, x_N^i]$, which all belong to the interval $[0, 1]$
2. for the ith uncontrollable factor define a mean u_i and a standard deviation σ_i corresponding to the area of interest
3. transform x_i by the transformation $\tilde{x}^i = [\Phi^{-1}(x_1^i), \ldots, \Phi^{-1}(x_N^i)]$ where $\Phi^{-1}()$ is the quantile function for the standard gaussian distribution
4. transform $\tilde{x}^i$ to $x_G^i = \mu_i + \sigma_i \tilde{x}^i$

This gives uncontrollable factor settings that independently of each other are gaussian with mean μ_i and standard deviation σ_i. Figure 5.6 illustrates the method for $N = 4 \times 25$ runs for one uncontrollable factor, which shows that the subdesigns can be assumed to be gaussian (p-values for shapiro-wilk's test for normality are $p > 0.93$ for the subdesigns and $p \approx 1$ for the combined design). This procedure can be generalized to other distributions by replacing $\Phi^{-1}()$ with the relevant quantile functions in step 3 and skipping or modifying step 4.

5.4 Papers D and E

Analysis of Computer Experiments with Multiple Noise Sources (European Network for Business and Industrial Statistics)

Paper D illustrates several modeling techniques for the output from simulating the surgical unit from section 2.2. The paper was expanded and modified to the journal article in Paper E and is hence covered by the summary for Paper E

Analysis of Computer Experiments with Multiple Noise Sources

Paper E is an extension of Paper D for the "ENBIS8" special issue in Quality Reliability Engineering International. The modeling techniques in Paper D are

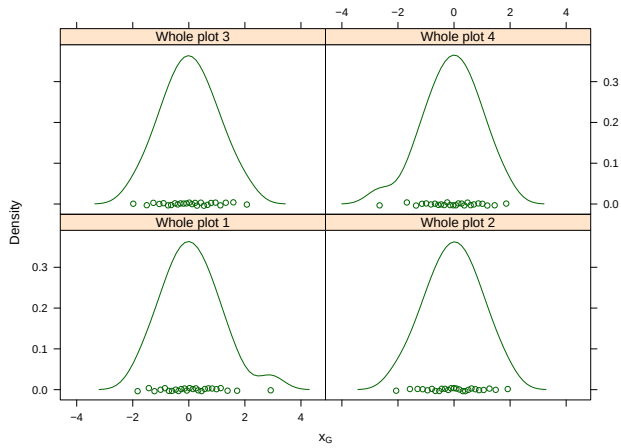


Figure 5.6: Transformation of uncontrollable factor settings in a top-down design to gaussian distributions

in this paper also evaluated on a simpler example in which the output function is known.

In Paper E we consider the M/M/m-queuing system to illustrate the methods applied on the more complicated model from section 2.2, since the M/M/m-queuing system is a well-known system and is expected to behave similarly to the simulation model. The M/M/m-queuing system has a vast literature and possesses many nice properties including that the expected waiting time is known (see section 2.3). This implies that the modeling techniques can be compared with the true underlying signal.

The simulation models considered are both influenced by uncontrollable factors and stochastic sources, which is dealt with in three different manners as described in section 4.2. The paper shows that the variation in the output can be split up in two sources by techniques known from physical experimentation. In a linear mixed effects model a variance component for the variation coming from changes in the settings of the uncontrollable factors and an estimator for the variance coming from changes in the seed (the random error) can be estimated. Moreover, the variation coming from changes in the setting of the uncontrollable factors can be analyzed and interpreted by means of generalized additive models (GAMs).

For the case-study two scenarios are considered: 1) the current setup and 2) 20 new settings of the controllable factor. The 20 new settings were found in

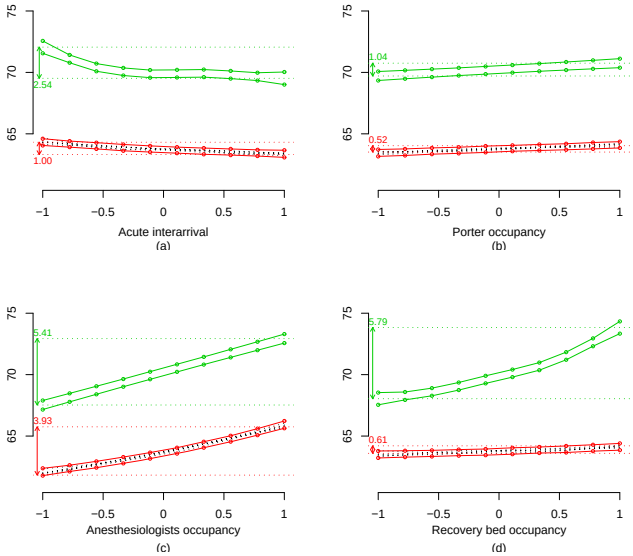


Figure 5.7: Estimated effects of the uncontrollable factors. Top curves are the reference setting and the bottom curves the new settings

a pilot study and were chosen such that the CVaR waiting time is expected to be low while maintaining the same total throughput and percentage of elective patients treated outside regular hours (EOUT). The analysis shows that with the current setting the output varies more both due to the uncontrollable factor settings and the seed, i.e., it is less robust compared to the new settings. This can also be seen from Figure 5.7, which shows that the estimated effects of the uncontrollable factors are flatter for the new settings compared to the current setup. Moreover, the estimated CVaR waiting time is 6.5 minutes shorter with the new settings, which shows that the improvement is significant. It was also shown that the methods worked well on the M/M/m-queuing system, i.e., was able to estimate the true function accurately.

The GAM framework also provides methods for handling binary and count outcomes, which in Paper E was used to estimate the likelihood that a new setting would perform at least as good as the current settings. The analysis highlighted three different settings of the controllable factors that had both higher throughputs, smaller percentages of elective patients treated outside regular hours and shorter CVaR waiting times compared to the current setting. All three settings suggested changing the number of operating days (for elective surgery) from five to four, i.e., fewer but longer days.

5.5 Paper F

2-stage approach for Kriging for simulation experiments with quantitative and qualitative factors

The topic of Paper F is Kriging for simulation models with quantitative and qualitative factors. The simulation model in section 2.2 is used for illustration of the extension of the Kriging interpolator after being put in a deterministic operating mode. The controllable factors are now thought of as being qualitative (they are ordinal having a few levels only), whereas the uncontrollable factors correspond to the quantitative factors. In section 4.1 the basic Kriging model is described and the following is based on those definitions. To ease the notation we denote one setting of the qualitative factors a whole plot, which reflects the structure of the top-down experiment (Dehlendorff et al., 2011) applied to the simulation model.

The usual correlation function given in equation (4.2) is now modified by including an extra term depending on the whole plots of the observations, i.e., $\tilde{R}(x_{ij}, x_{kl}) = R(x_{ij}, x_{kl}) \cdot (I(i = k) + I(i \neq k)\alpha_{ik})$, where x_{ij} is the ith whole plot and jth observation. Five different correlation structures are considered

1. $\alpha_{ik} = \theta_c$: correlations between observations from different whole plots are reduced by a constant quantity
2. $\alpha_{ik} = g(\hat{\mu}_i, \hat{\sigma}_i, \hat{\mu}_k, \hat{\sigma}_k)$: correlations between observations from different whole plots are reduced by a quantity depending on the sample means and standard deviations of whole plot i and k
3. 2-stage procedure (described below)
4. $\alpha_{ik} = \exp\left(-\sum_{q=1}^{d_z} \theta_{zq} I(z_i^q \neq z_k^q)\right)$ where z_i^q is the level of the qth qualitative factor for the ith observation (see Hung et al. (2009))
5. α_{ik} is parameterized by a hypersphere parameterization as proposed by Zhou et al. (2010)

In the 2-stage procedure we first fit a Kriging model for each whole plot in the quantitative factors

$$Y_i(\mathbf{x}_{ij}) = \mu_i + Z_i(\mathbf{x}_{ij}) \quad i = 1, \ldots, m \tag{5.2}$$

New design sites are then generated by adding the correlation parameters of model i to all observations from whole plot i, i.e., the design sites becomes

$\tilde{\mathbf{X}} = \left[\mathbf{X} \quad \left[\mathbf{C}_1^T \otimes \mathbf{1}_{1\times q_1} \quad \dots \quad \mathbf{C}_m^T \otimes \mathbf{1}_{1\times q_m}\right]^T\right]$ where $\mathbf{X}$ is the original design sites ordered by whole plot, q_i is the number of observations from whole plot i and $\mathbf{C}_i$ the correlation parameters for whole plot i. An overall model is then estimated using $\mathbf{y}$ and the new design sites $\tilde{\mathbf{X}}$ using the standard Kriging model in (4.1). This implies that whole plots that have similar correlation structure are defined to be close and therefore correlated. The idea is similar in the mean-variance case, but now $\mathbf{C}_i = [\hat{\mu}_i \ \hat{\sigma}_i]$. This structure assumes that whole plots with the same mean and variance are similar.

The five correlation functions are first evaluated on six test functions, which shows that the mean-standard deviation and 2-stage procedures give the most accurate meta-models. On two realistic examples using the simulation model of the surgical unit from section 2.2 the 2-stage procedure outperforms the other correlation structures.

One drawback of correlation structures 1-4 is that they can not handle negative correlations between whole plots, which is possible with the fifth correlation structure. However, the flexibility of the fifth correlation structure comes with a price, which is the number of parameter required for correlation between whole plots. This may result in overfitting for small data sets with many levels of the qualitative factors, which is a likely scenario since computer and simulation models tend to be very time consuming and have many factors. In the 2-stage procedure several Kriging models are fitted, they are however somewhat easier to fit since they are fitted on subsets of the data set in the initial step. Moreover, the total number of correlation parameters in the combined model is twice the number of quantitative factors and thus still manageable.

Kriging is a very powerful tool and many new methods within simulation are based on this method. Stochastic Kriging models as considered by van Beers and Kleijnen (2008) and Ankenman et al. (2010) handle simulation models with stochastic output. Robustness analysis through Kriging is also a relatively new topic and is for example considered by Dellino et al. (2009), who fit separate Kriging models for the mean and standard deviation to estimate the Pareto frontier. The method in this paper is seen to perform well on a simple yet realistic case-study and hence is an alternative the more complex model by Zhou et al. (2010).

Chapter 6

Discussion

Design and analysis of computer and simulation experiments is a relatively new research area. Many challenges are encountered in this area and hence a wide range of methods has been developed. In this thesis contributions in both the design and the analysis part of the area are introduced.

The first major contribution is the development of the top-down experiment, which provides an experimental plan with a better coverage of the uncontrollable factor space compared to the crossed design. Furthermore, application of the design on a simulation model showed that the coverage of the uncontrollable factors improved the understanding of the interactions between controllable and uncontrollable factors. The design is based on uniform designs and one idea for future research is to consider different underlying designs such as, e.g., the maximin design.

Qian et al. (2009a) and Qian et al. (2009b) consider nested space-filling designs, i.e., a high accuracy experiment is nested within a low accuracy experiment. Qian and Wu (2009) consider sliced space-filling designs based on orthogonal designs. The overall design principal of the nested and sliced space-filling designs are seen to be similar to ours, i.e., that the design on both the overall and sub level is taking to account. In future research it would be interesting to compare the performance of the top-down design with the nested and sliced designs. Another interesting approach is sequential sampling as considered by Kleijnen

and van Beers (2004) and van Beers and Kleijnen (2008), who use an adaptive sampling scheme, i.e., the next sampling point is based on a criteria based on the information from the already simulated settings. The adaptive procedure may serve as a benchmark for evaluating the performance of deterministic sequential sampling based on the top-down design structure as discussed in section 5.3.

The second area of contribution is related to output analysis of simulation models. First the CVaR statistic for waiting time distribution was introduced. Next methods for analyzing simulation models with multiple noise sources were considered, and finally a method for Kriging for analyzing computer and simulation models with quantitative and qualitative factors was proposed.

The CVaR statistic is a measure originating from finance as a measure of risk. CVaR is relevant if the long waiting times are the primary concern, whereas the average waiting time may be more appealing to the management for example if the waiting times are related to the staff and not the patients. One drawback of the CVaR criteria is that the required size of the sample increases as $(1-\alpha)$ decreases. However, it may be seen as a robustness measure, i.e., a low CVaR (close to the mean) indicates a setting that is robust since it implies that the risk of long waiting times is low.

For stochastic simulation several modeling techniques from physical experimentation were considered, which were shown to perform well for our case-study. Stochastic Kriging is introduced in a recent paper by Ankenman et al. (2010), who include an extra stochastic element in the usual Kriging model to account for the variation from one replicate to the next. Kriging is a very flexible and powerful meta-model for deterministic simulation and hence the stochastic version is expected to be useful in applications, in which for example regression methods fail. Fitting Kriging models for the average at each setting is another method to deal with stochastic simulation as considered by van Beers and Kleijnen (2003) and Kleijnen (2008), who apply boot-strapping to estimate the uncertainty related to the replications.

Finally a Kriging model for simulation models with quantitative and qualitative factor is introduced. The fitting procedure is done in two steps and each step consists of ordinary Kriging models with simple correlation structures. Zhou et al. (2010) also consider Kriging for models with quantitative and qualitative factors and introduces a parameterization that can handle negative correlation between different settings of the qualitative factors, which is not handled in our method. For a simple yet realistic case-study it was shown that our method performed better than the method by Zhou et al. (2010), it is however expected that their method will perform better in cases where negative correlations are present. Moreover, if the number of qualitative factors is low and the number of quantitative factors is high the model by Zhou et al. (2010) uses fewer param-

eters compared to our method, whereas with many qualitative factor settings our method is more efficient in terms of the number of parameters.

Kriging is a popular method and interesting extensions to the Kriging model may be analysis of models with multiple outputs and robustness studies as considered by Dellino et al. (2009). In this thesis several methods for analysis of the output from our case-study have been considered and robustness is an interesting extension of our current results. Our results based on regression methods indicate that the case-study may be put in a more robust operating mode, but using methods based on Kriging may expand the knowledge about the uncontrollable factors.

Paper A

Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units

---

Accepted for publication in Quality Technology and Quantitative Management, Volume 7(2) September 2010, p. 321-336

Conditional Value at Risk as a Measure for Waiting Time in Simulations of Hospital Units

Christian Dehlendorff[1*] Murat Kulahci[1] Søren Merser[2]

Klaus Kaae Andersen[1]

[1]DTU Informatics

Technical University of Denmark

[2]Clinic of Orthopaedic Surgery

Frederiksberg Hospital

Abstract

The utility of conditional value at risk (*CVaR*) of a sample of waiting times as a measure for reducing long waiting times is evaluated with special focus on patient waiting times in a hospital. *CVaR* is the average of the longest waiting times, i.e. a measure at the tail of the waiting time distribution. The presented results are based on a discrete event simulation (DES) model of an orthopedic surgical unit at a university hospital in Denmark. Our

*cd@imm.dtu.dk

analysis shows that *CVaR* offers a highly reliable performance measure. The measure targets the longest waiting times and these are generally accepted to be the most problematic from the points of view of both the patients and the management. Moreover, *CVaR* can be seen as a compromise between the well known measures: average waiting time and the maximum waiting time.

Keywords: *Waiting time distribution, Conditional Value at Risk, Simulation, Health Care*

1 Introduction

Simulation studies are widely used in health care applications due to the large number of uncertainties involved. The complexity of these systems together with the physical and legal constraints in the actual systems make simulation a very powerful tool for experimentation to serve as a basis for analytic optimization methods [4, 9].

Simulation models in health care applications are used both for optimization of existing facilities [8] and in planning new facilities [18]. Ferrin and McBroom [8] maximized hospital revenue by process improvements in the emergency departments. Length of stay (*LOS*), the number of patients leaving without receiving care, the percentage of admissions accepted and ambulance diversion hours were used as outcomes. Miller et al. [18] considered the merging of six emergency departments into one and focused on the average *LOS*. Their results show that the *LOS* can indeed be considerably reduced. They further show that the distribution of *LOS* is right-skewed with a long tail. Jun et al. [14] reviewed the health care simulation literature and concluded that simulation is often used to optimize allocations and as a tool in staff planning. They cited various studies related to patient scheduling and to staff sizing and planning. They also reported that many studies use trade-offs between the utilization of doctors, rooms etc. and patients' waiting times as outcomes.

Denton et al. [7] studied expected surgical suite waiting time, surgical suite idle time and total overtime and used a linear trade-off combination of these mea-

sures as a single measure. This linear combination is a cost measure which takes into account the discomfort of patient waiting time and considers it together with the lost revenue corresponding to idle surgical suite time and the cost of overtime.

Cayirli and Veral [5] reviewed out-patient scheduling and summarized a number of possible performance measures related to the quality of such systems. The time-based measures included the mean, the maximum and the frequency distribution of the waiting times. Their summary for the suggested performance measures showed that the majority of studies used mean waiting time, total costs of waiting, percentage of patient waiting less than a certain threshold, and the variation of waiting time.

The main objective in this article is to compare Conditional Value at Risk (*CVaR*) as a optimization measure for patients' waiting time with existing measures and to report on the performance of this new measure based on a specific case-study of an orthopedic surgical unit. The concept of *CVaR* is formally introduced in section 3.1 and originates from economics. *CVaR* was introduced by Rockafellar and Uryasev [21] as a measure to quantify a distribution of losses; typically in portfolio scenarios. The measure was introduced as an extension to Value at Risk (*VaR*), one of the most commonly used performance measures in portfolio management. The *CVaR* criterion focuses on the right tail of the loss distribution and provides a measure of the expected value of the highest losses. The *CVaR* criterion has been used in a wide variety of applications (see for example [1], [10] and [27]), but not in the context of our study. The suggested use of *CVaR* is for optimization of a given system's performance in terms of waiting time

and is relevant in cases where the frequency of long waiting times is the primary concern.

In this article, a discrete event simulation model of an orthopedic surgical unit in Copenhagen, Denmark is presented as the case-study. The long term goal for the simulation study is to minimize the total waiting time, with special focus on long delays. In the case-study analysis of the uncertainties and behaviour of different performance measures including *CVaR* under various resource and simulation settings are presented. Moreover, *CVaR* is compared to other measures using this model as illustration. The article is structured in the following way: Section 2 describes the case-study. *CVaR* is defined in section 3 followed by section 4 where the performance measure is evaluated by considering the simulation model under different resource and simulation setups. Finally the key findings are summarized in section 5.

2 Simulation model

In this section, we present our case-study for evaluating the performance of the *CVaR* waiting time criterion in the simulation of an orthopaedic surgery unit. The level of detail of the model is intentionally kept low, since our main objective is to use it as an illustration of the *CVaR* measure.

2.1 The surgical unit

As in much of the rest of the world, over the past decade the Danish public health care system has been subject to increasing demands for efficiency [14]. The system is now under considerable pressure for higher throughput in order to reduce waiting lists. Avoiding or reducing delays in the system is certainly one of the many options to reach this goal. Furthermore, fewer and/or shorter delays may also increase patient satisfaction, an issue that is central to today's quality and productivity improvement strategies in general.

The case-study is a surgical unit, which is part of an orthopedic department at a university hospital in Copenhagen, Denmark. The unit undertakes both acute and elective surgery and performs more than 4,600 operative procedures a year. While the patients come from various wards throughout the hospital, the main sources of incoming patients are the four stationary orthopaedic wards or the emergency care unit. The outpatients treated in outpatient clinics are not considered in this model but the resources shared between outpatients clinics and the surgical unit are included. Also day-case surgery patients with short recovery times are included in the model.

2.2 Model description

The conceptual model is outlined in Figure 1. It consists of three main modules: 1) the incoming module with arrival and wards, 2) the surgical unit with preparation and operating rooms and 3) the recovery. Module 3 is linked back to module 1,

since the patients return to the wards for final recovery and discharge.

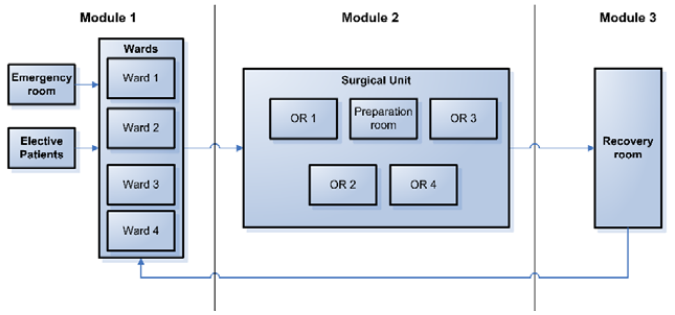


Figure 1: Conceptual model for an orthopedic surgery unit. The 3 modules are separated by vertical lines and the arrows indicate the patient flow

The simulation model is implemented in Extend^{TM} version 6 [17] and controlled from a Microsoft Excel spreadsheet with a Visual Basic for application script. The patient flow is outlined in Figure 2. All patients are either acute or elective and are admitted to one of the four stationary wards from where the patients are collected when an operating theater is ready. Patients are then either sedated, sent to a preparation room and brought to the operating room or brought directly to the operating room for sedation and preparation. The patients are operated and hereafter attended to by an anesthesiologist before being moved to the recovery room. As the patients are moved out of the operating room, cleaning and preparation of the rooms for the next patients are started.

The resource constraints in the system are process related: available surgeons for the operation, a free recovery bed and an available porter for moving the patient to the recovery room, etc. These resources are controlled by a central mechanism

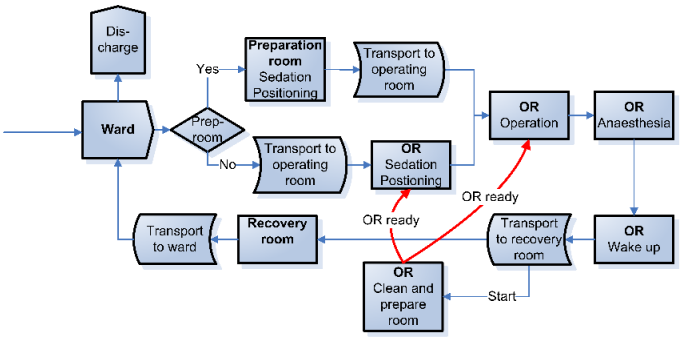


Figure 2: Process diagram for patient flow through the system from ward to discharge.

controlled by different schedules, e.g. more resources during regular hours. Sharing between different specialties is handled with the resource pools. In our model the resources include staff and physical facilities such as operating rooms and recovery beds. It should be noted that some resources such as surgeons, anesthesiologists, porters and recovery beds are shared with other departments or procedures not directly related to the surgical unit.

2.3 Empirical Data

Prior to the simulation study, a simple registration of the time from patients' arrival at the surgical unit until their departure to the recovery room was done by the staff for a period of 3 months. The initial data set held no information on subprocesses, which implied that a more elaborate registration system was needed. In the new registration system, the nurses at the surgical unit recorded the patient flow through the unit from the ward to the recovery room, i.e. each subprocess

was recorded over a period of 1 month.

The new data was validated on the data collected routinely by the staff prior to the simulation study by comparing the total time spent at the surgical unit recorded in the two data sets with a Kolmogorov-Smirnoff (K-S) goodness of fit test [6], which indicated no significant difference. Furthermore, tests for correlation [12, 2] between processes in the new data set indicated that the subprocess durations were statistically uncorrelated indicating that subprocesses could be modeled individually.

2.4 Validation and verification

The model was inspected graphically by the management of the department to verify the patient routing and the procedures. Animation was included in the model to assist and simplify verification during the presentation of the model.

Model validation corresponding to patient volume and waiting time was carried out by comparing the simulation output with the observed data. All validation was carried out using graphical methods (QQ-plots, density plots and histograms) and formal statistical tests (K-S and Wilcoxon rank-sum tests [13]) with a significance level of 5 %. A more elaborate validation was also carried out corresponding to the scheme outlined by Sargent [22] and although this concluded that the model was adequate, it is not presented in this article.

The model parameters were calibrated on the individual processes and queuing times, and finally validated on the total duration defined as the time from the patient leaving the ward to the the time the patient is moved to the recovery

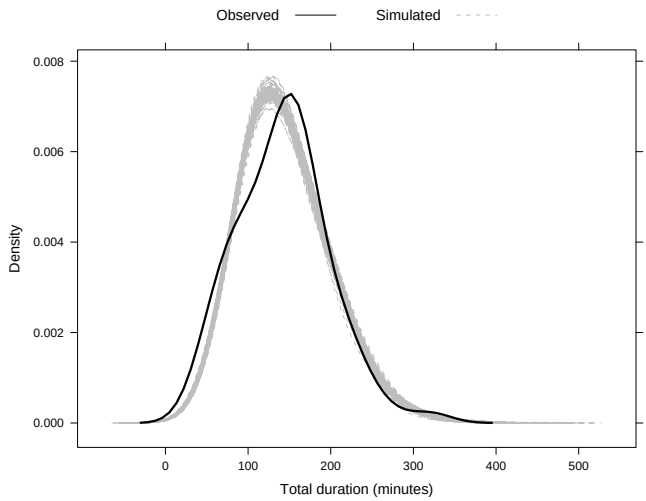


Figure 3: Estimated density functions for observed (black solid line) and 100 simulation runs (gray dotted lines) for total time at surgical unit

room. Figure 3 shows that the model tends to mimic the system's overall behavior, which was confirmed with K-S and Wilcoxon tests indicating no statistical difference. The throughput, mixture of patients and distribution of patients per day were validated as a part of the tuning and calibration process.

The incoming rate of elective patients per day was shown to fit a discretized triangular distribution function, which was also validated by a K-S test. The acute patients were assumed to have exponentially distributed inter-arrival times. K-S tests indicated that the distribution of acute patients per day and the ratio of elective to acute patients were modeled adequately. The acute incoming rate was much more volatile compared to the one for elective patients. The coefficient of

variation (*CV*), which is defined as the standard deviation divided by the mean, was 2.5 times higher for the acute patients compared to the elective patients. In both cases the variation in the observed data set was large with *CV* greater than 90 %.

3 Performance measures

One of the most essential issues in any simulation study is to define sound and reliable performance measures [19]. Each simulation run is summarized in a set of measures, which characterizes the overall performance of the system. Often more than one measure is investigated in order to quantify the objectives of the study, e.g. avoiding long waiting times while keeping a certain level of patient throughput. In this paper Conditional Value at Risk is introduced as a waiting time measure targeting the longest waiting times and compared to other existing measures.

3.1 Conditional Value at Risk

Conditional Value at Risk is a concept originating from finance as a measure of risk [21, 15, 16]. For a distribution of waiting times, T, CVaR is defined as the expected value of the $(1-\alpha)$-tail of T, i.e. given as

$$CVaR_\alpha(T) = E[T|T > q_\alpha] \tag{1}$$

where q_α is the α-quantile, where $P(T \leq q_\alpha) = \alpha$. For a sample of simulated waiting times, $T_x = \{t_{x1}, \ldots, t_{xN}\}$ (obtained from the x^{th} run), the $CVaR_\alpha(T_x)$ is estimated by

$$CVaR_\alpha(T_x) = \frac{1}{1-\alpha}\left[\left(\frac{i_\alpha}{N} - \alpha\right) t_{xi_\alpha} + \sum_{i=i_\alpha+1}^{N} \frac{t_{xi}}{N}\right] \tag{2}$$

with $t_{x1} \leq t_{x2} \leq \cdots \leq t_{xN}$, i_α is the index satisfying $\frac{i_\alpha}{N} \geq \alpha > \frac{i_\alpha - 1}{N}$, t_{xi_α} is the α-quantile and in economics denoted as the Value at Risk (*VaR*). *VaR* is seen to be indifferent to the shape of the $(1-\alpha)$-tail, i.e. a given *VaR* value covers situations from short $(1-\alpha)$-tails to long $(1-\alpha)$-tails. In most applications of *CVaR* the estimate is based on the $(1-\alpha)100\% = 5\%$ longest waiting times and in the following CVaR is therefore estimated by equation (2) with $\alpha = 0.95$.

For waiting times the *VaR* waiting time is the value of the α-quantile of the total waiting times, e.g. for $\alpha = 0.95$, 95 % of the patients have a total waiting time less than or equal to *VaR*. *CVaR* is the average of the 5% longest waiting times, i.e. a measure about the tail of the waiting time distribution. It is seen that *CVaR* is at least as large as *VaR* and that the difference indicates the skewness of the distribution, hence the two measures are correlated. *CVaR* is seen to be more sensitive to samples with very long waiting times compared to *VaR*. However, Webby et al. [27] noted that *CVaR*, as opposed to *VaR*, is more stable with changes in the α-value. This can be explained by the fact that *CVaR* is an average of the tail, whereas *VaR* is the quantile defining the tail. The quantile is likely to jump with a small sample, whereas the average will shrink this effect.

The rationale for introducing *CVaR* waiting time measure is that it is a well known measure of risk in finance. It fits well in an optimization framework with the objective of minimizing the overall waiting time while controlling the risk of experiencing very long waiting times. The tail of the waiting time distribution in these studies is quite important since as shown by Bielen and Demoulin [3], in terms of patient satisfaction, waiting time influences satisfaction negatively. That is, longer waiting times decrease patient satisfaction significantly. Using the average waiting time inherently imply that the distribution of the waiting times is unimportant as long as the overall waiting time is low. This is, however, not in accordance with patient satisfaction and quality perception. On the other hand the maximum waiting time may be a too risk averse measure and could potentially confound good settings with bad settings since it is based on only the most extreme observation.

The benefits of using *CVaR* as performance measure are that it is easy to compute, easy to interpret and targets the long waiting times. As mentioned above, if the mean waiting time (denoted risk neutral) is used, an increase in the longest waiting times can be overlooked since a shift in the tail may be averaged out by the rest of the distribution. On the other hand, using the maximum waiting time (risk averse) may corrupt the results, since a single long waiting time may be an outlier in an otherwise well performing setup. *CVaR* can be seen as a compromise between the average waiting time ($\alpha = 0$) and the maximum waiting time ($1 - 1/N < \alpha < 1$), with $(1-\alpha)$ reflecting the risk of long waiting times. Hence a low α corresponds to a high risk of overlooking long waiting times since the

importance of these is low.

3.2 Other measures

Other measures have been suggested in the health care literature, which are discussed in the following. Tang et al. [26] presented mean residual life, i.e. the expected residual life time given that a unit has lived a certain amount of time. In terms of waiting time this is equivalent to the expected residual waiting time having waited a certain amount of time. Length of additional stay (*LAS*) is another metric for measuring waiting times, Silber et al. [24] defined it as the remaining length of stay (*LOS*) after the transition point at which the stay becomes prolonged. A stay may become prolonged at the first time point, x, where the probability of a total length of stay of $x+y$ is greater than the probability for a *LOS* of y from the beginning. The test for the prolonging point is done with the Hollander-Proschan test [11]. *LAS* is seen to be the mean residual life at the point where the stay becomes prolonged. The rationale behind *LAS* is that if a stay is prolonged it is more likely to be associated with a complicated case [24].

Both *LAS* (the *MRL* at the prolongation point) and *MRL* are similar to the *CVaR* measure. However, *CVaR* is the expected waiting time of the $(1-\alpha)100\%$ longest waiting times, whereas mean residual life at the α-quantile is the expected remaining waiting time after having waited t_{xi_α} minutes. Silber et al. [24] suggest using the point at which a stay becomes prolonged as the choice for α. This imply that for different setups the corresponding *LAS*'s (or *MRL*'s) are the average residual waiting times for the prolonged stays, i.e. for different α-values. Fur-

thermore, the scale is different depending on the setting: in one case it may the residual waiting time after having waited 30 minutes while in another it may be the residual waiting time after having waited 60 minutes. For *LAS* and *MRL* in general unlike for *CVaR* the interpretation is seen to be dependent on the distribution. This implies that the scale and interpretation are maintained for different settings, which makes it suited for use in optimization. Moreover, the distribution of waiting times may be on time, i.e. no prolongation point is present, which implies that the *LAS* concept breaks down.

From a quality point of view the waiting time may be more interesting than the residual waiting time, since the patient's perception of the quality of the treatment is related to his/her total waiting time and not the residual waiting time after having already waited for x minutes. In terms of waiting times the length of additional stay may not be as important as for the length of a hospital stay, since the waiting time indicates something about the system's performance and not of the severity of the operation or complications for the individual patient. Moreover, the waiting time is the time between activities and hence complicated cases have longer activity times and more difficult recovery, which do not influence the waiting time. Silber et al. use the *LAS* as an indicator of health care outcomes and the measure is hence not targeted at evaluating a system's performance. The *LAS* framework does not seem to be well suited for evaluating waiting times, whereas it is highly relevant for seeking complicated hospital stays.

4 Case Study

This section presents the performance measures by applying them to output from the simulation model presented in section 2. The measures are initially examined under the existing setup in terms of the variation from run to run and the sensitivity to length and number of runs. They are then considered under different resource settings. The proposed measure, *CVaR*, is analyzed and compared to other well known measures presented in section 4.3.

4.1 Simulation setup

The simulation model is run for at least 300,000 minutes (see section 4.4). This corresponds to 30 weeks with a warm-up period of 10,080 minutes (1 week) for each run. In each run different performance measures are obtained as described in section 4.3. These measures are summarized by their minimum, maximum, average and coefficient of variation (sample standard deviation in % of the average) across runs.

4.2 Analysis methods

The results from the simulation model are analyzed using statistical test methods. Wilcoxon two-sample tests [13] are used to compare two samples in terms of their location. The test is a non-parametric test. Comparing two samples in terms of their distributions is done with Kolmogorov-Smirnoff two- sample test [6], which is also a non-parametric method. Here we compare the empirical distributions

and test whether they can be assumed to be identical. Significance of correlation coefficients is tested based on Spearman's rho [12, 2], a non-parametric approach based on ranks. The main rationale for using non-parametric tests is that they do not rely on specific distribution assumptions and are robust against outliers. All data analysis was done in R version 2.7.1 [20].

Densities functions are estimated with the `density` procedure from the `stats`-package and plotted with the `densityplot` function from the `lattice`-package in R [20, 23] using the default values. The defaults are a Gaussian kernel with a bandwidth, $h = 0.9n^{-1/5}\min[\hat{\sigma}_x, IQR_x/1.34]$, where x is the sample, which has sample standard deviation $\hat{\sigma}_x$, inter-quartile range IQR_x and sample size n (Silverman's rule-of-thumb) [25].

4.3 Performance measures

The main focus of the simulation study is on the waiting times defined as the time wasted between processes and is measured in minutes. For each patient a number of waiting times are identified: waiting time before the surgeon talks to the patient before sedation, waiting time for the anesthesiologist, waiting time before operating room is ready and waiting time for a porter and a free recovery bed, etc. The total waiting time for the j^{th} patient in the i^{th} simulation run, t_{ij}, is estimated as the sum of K sub waiting times, t_{ijk}. The waiting time measures considered in this article are

- Average waiting time, $\overline{WT}$

- Maximum waiting time, MWT
- Conditional Value at Risk, *CVaR*, waiting time, $CVaR(WT)$
- Value at Risk, *VaR*, waiting time $VaR(WT)$

Additionally total throughput (total number of patients treated, TT) and percentage of elective patients treated outside regular hours, $EOUT$, are considered. These measures are included in the simulation study to ensure that the throughput remains the same and the elective patients are not treated outside regular hours, hence without creating additional costs due to overtime. The average and maximum waiting times are frequently used measures to quantify the waiting time [5]. *VaR* is included to highlight the additional information contained in our main measure, *CVaR*, and to illustrate its close relationship to *CVaR*.

4.4 Run length and sample size analysis

The first example consists of simulations on the system at its current configuration. Here, the main objective is to examine the performance measures under different run lengths and numbers of repetitions (runs). Table 1 shows the summary for three types of simulation runs for the system as it is: 1) 30-weeks simulation repeated over 100 runs, 2) 30-weeks simulation repeated over 200 runs and 3) 60-weeks simulation repeated over 60 runs.

From the first block in Table 1 it is seen that the total waiting times are highly skewed with an average $\overline{WT}$ of around 31 minutes, a 95 % quantile of around 61 and a maximum of 111 minutes. It is seen from the CV column in the first block

Table 1: Summary for performance measures over runs, e.g. the minimum, maximum, average and *CV* of total throughput for three types of simulation setups. The *Min*-entry for the first row e.g. summarizes the minimum $\overline{WT}$ of the 100 runs, *Max* the maximum, *Mean* the average and *CV* the standard deviation in percent of the mean. The units for the waiting time statistics are minutes, the unit for EOUT is percent and TT is measured in number of patients.

	Min	Max	Mean	*CV*(%)
	30 weeks, 100 runs, 3 porters			
$\overline{WT}$	30.03	32.21	30.97	1.52
MWT	89.00	157.88	111.25	11.34
TT	1635	1797	1711	2.02
$EOUT$	8.25	12.69	10.15	9.22
CVaR	67.98	77.47	71.17	2.26
VaR	58.05	64.01	60.95	1.92
	30 weeks, 200 runs, 3 porters			
$\overline{WT}$	29.69	32.29	30.98	1.49
MWT	89.00	163.36	111.92	11.48
TT	1615	1827	1715	2.15
$EOUT$	8.25	12.97	10.36	9.15
CVaR	67.58	78.09	71.36	2.30
VaR	58.05	64.40	60.94	1.99
	60 weeks, 60 runs, 3 porters			
$\overline{WT}$	30.21	31.52	30.91	0.94
MWT	94.30	153.97	118.57	10.27
TT	3347	3599	3468	1.82
$EOUT$	8.95	11.73	10.51	5.91
CVaR	67.90	73.35	71.17	1.43
VaR	58.96	62.16	60.69	1.25

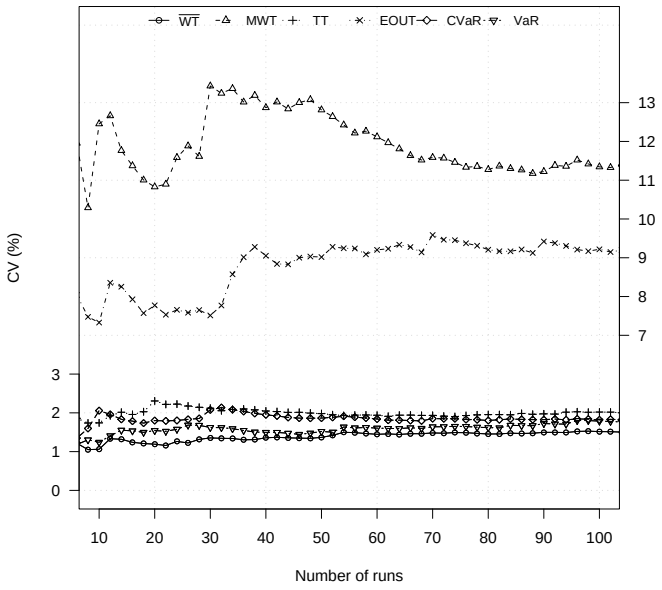


Figure 4: Coefficient of variation as function of included runs for the 6 performance measures

in Table 1 that the most varying measure is the MWT ($CV = 11.3\%$) followed by $EOUT$ ($CV = 9.2$ %). The remaining four measures are comparable in terms of coefficient of variation ($1.5\% \leq CV \leq 2.5\%$).

Figure 4 illustrates the evolution of the CV's as the number of runs is increased. It can be seen that all CV's are stabilized after 70 runs, however subdivided into the two groups as described previously. It can also be seen that the two upper curves take more runs to settle in compared to the bottom four. Clearly the maximum

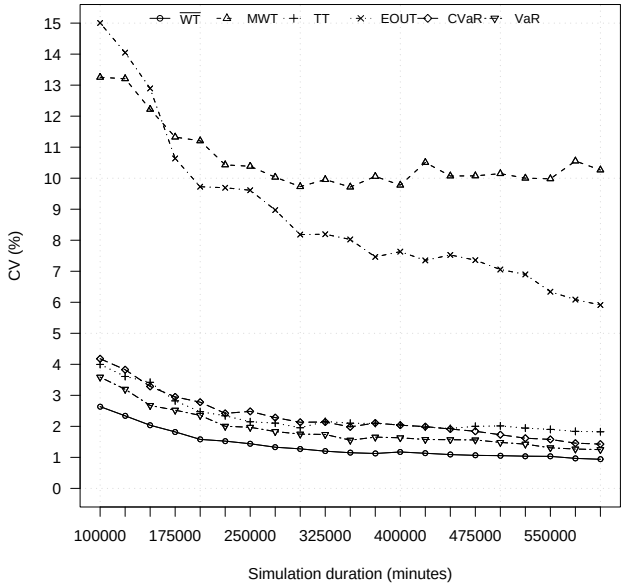


Figure 5: Coefficients of variation for 60 runs with varying run lengths for the 6 performance measures

waiting time is a measure highly dependent on the simulation run, since it is the most extreme observation in each run. The average waiting time is as expected the least varying measure, whereas the *CVaR* and *VaR* are seen to vary almost equally much. Figure 4 indicates that the four best performing measures have stabilized after 30-40 repetitions.

Figure 5 shows that a run-length of 300,000 minutes (30 weeks) seems to be adequate for obtaining a low *CV* for 5 out of 6 measures (no significant im-

provements hereafter). $EOUT$ is seen to be improving by more than 2 %-points from 300,000 minutes to 600,000. Simulating 30 weeks repeated 60 times is a good trade-off between simulation time and precision for MWT, which leads to an approximate half width of a 95 % confidence interval for the average of MWT corresponding to 2.7% of its estimated value. For $\overline{WT}$, TT, *VaR* and *CVaR* considerably fewer repetitions are needed. In fact Figure 4 suggests that fewer than 20 repetitions will be sufficient.

In the 100 run simulation of 30 weeks each *CVaR* is significantly correlated with *VaR* (as expected), MWT and $\overline{WT}$. Moreover, *VaR* is significantly correlated with $\overline{WT}$, whereas TT is correlated both with $EOUT$ and $\overline{WT}$. The correlations are all positive, which implies that higher throughput is associated with longer waiting times. The *VaR* is seen to be uncorrelated with the MWT, whereas *CVaR* is. This in fact fits well with the definition of *CVaR* and *VaR*. The connection between *CVaR* and $\overline{WT}$ and MWT was shown in section 3.1.

4.5 Sensitivity to Changes in Resource Allocation

The sensitivities of the measures to changes in resource allocation are analyzed by changing the number of porters at the surgical unit in regular hours. Three porters are available in regular hours in the current system described in section 4.4. This number is set to 1, 2 and 4 in the following analysis. The porters are a relatively less costly resource to adjust than the number of surgeons, nurses and operation rooms. The expectations are that lowering the number of porters will increase the waiting times and decrease the throughput or increase the percentage of patients

being treated outside regular hours. Hence increasing the number of porters may enable an increase in the incoming flow of patients without increasing the waiting times if the remaining resources are underutilized in the current setup.

Table 2 summarizes the results from 60 runs of 30 weeks for three different settings of porters. It can be seen that having 2 or 4 porters are equivalent with the results for 3 porters in Table 1, whereas having 1 porter clearly increases the waiting times in terms of the average, *CVaR* and *VaR* waiting time. In the top part of Figure 6 the associated estimated density functions indicate that 2-4 porters lead to equivalent waiting time distributions, whereas the 1 porter distribution seems to differ.

With 1 porter it is observed that all measures besides the total throughput are changed significantly (Wilcoxon two-sample test [13]: $p < 0.001$) compared to having 3 porters. The patients wait longer on average (8.56 % increase on average), have higher maximum waiting times (8.41 % increase on average), more patients are treated outside regular hours (19.41 % increase on average) and *CVaR* and *VaR* are increased significantly (7.53 % and 6.97 %, respectively). Figure 6 shows that the primary change from 2-4 porters to 1 porter is a heavier tail. This is reflected in the *CVaR* in Table 1 and 2, which show that the increase is around 2 times the increase in the average waiting time. The top part of Figure 6 shows that the estimated density function with 1 porter is flatter around the peak and has a thicker tail, which increase the *CVaR* more than $\overline{WT}$. The increase by 5 minutes in CVaR from 3 to 1 porter corresponds to an increase in waiting time for the approximately 85 patients with the 5 % longest waiting times of 7 hours. In our

Table 2: Summary for performance measures over runs for three different configurations as in Table 1. The units for the waiting time statistics are minutes, the unit for EOUT is percent and TT is measured in number of patients.

	Min	Max	Mean	CV(%)
	30 weeks, 60 runs, 4 porters			
$\overline{WT}$	29.85	31.93	30.89	1.56
MWT	92.58	161.64	113.01	12.17
TT	1609	1812	1710	2.45
$EOUT$	6.87	12.69	10.48	9.99
$CVaR$	66.97	74.24	71.17	2.44
VaR	58.09	63.35	60.67	1.97
	30 weeks, 60 runs, 2 porters			
$\overline{WT}$	30.08	32.34	31.16	1.42
MWT	87.70	139.94	110.49	10.27
TT	1629	1815	1718	2.42
$EOUT$	8.38	13.05	10.88	8.11
$CVaR$	67.79	75.33	71.13	2.54
VaR	57.99	63.14	60.92	1.89
	30 weeks, 60 runs, 1 porter			
$\overline{WT}$	32.70	34.42	33.62	1.16
MWT	97.88	151.27	120.01	10.86
TT	1625	1815	1715	2.41
$EOUT$	10.36	14.08	12.12	6.51
$CVaR$	71.78	80.61	76.53	2.37
VaR	62.40	67.65	65.20	1.92

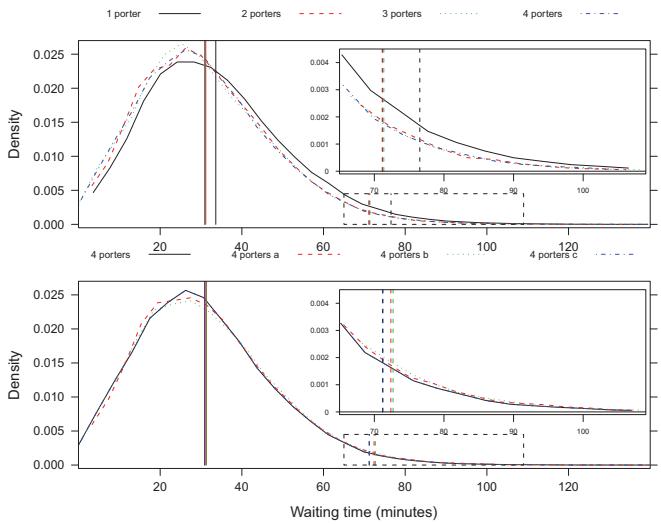


Figure 6: Estimated densities for 1, 2, 3 and 4 porters (top) and 4 porters with different patient load (bottom). Vertical lines correspond to mean waiting time (solid lines) and *CVaR* value (dashed lines). Porters 4a, 4b, and 4c correspond to 4 porters with 7 %, 14 % and 29 % more elective patients, respectively. The dashed area in the lower right of each panel is highlighted in the upper right.

simulation study the difference in CVaR is statistical significant, but the practical importance of the increase may be limited.

Adding an extra porter does not shorten the waiting times (top block in Table 2), the situation is comparable with the original 3 porter setting. The performance measures were not significantly different. The lowest p-value is obtained for *VaR* with a p-value of 0.18. Figure 7 furthermore shows that increasing the number of elective patients leads to a significantly worse performance compared to both the 3 and 4 porter situation (for all measures other than MWT). The bot-

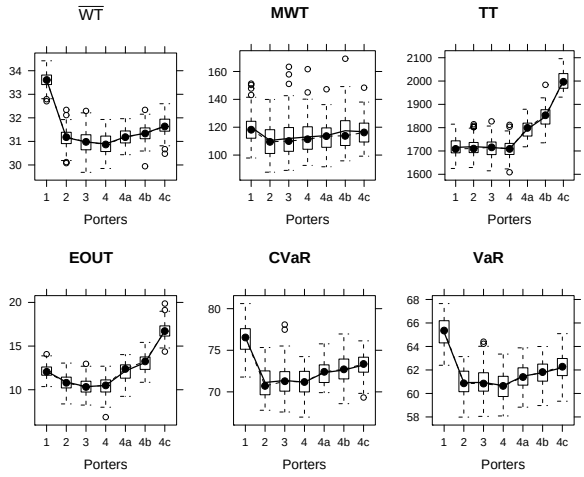


Figure 7: Box plots for comparing performance criteria for different resource settings. 4a, 4b, and 4c correspond to 4 porters with 7 %, 14 % and 29 % more elective patients, respectively.

tom part of Figure 6 indicates that the patients are waiting longer on average as the incoming rate is increased and that the tail of the waiting time distribution has the same length (MWT the same) but is heavier (*VaR* and *CVaR* increased).

It can be seen that *CVaR* has a higher absolute increase compared to $\overline{WT}$ for the 3 vs. 1 porter comparison, showing that the 5 % longest waiting times are increased the most. For increased patient input MWT does not increase, whereas *CVaR* and *VaR* do. This shows that using the MWT as criterion for judging the waiting time performance is a poor choice as it may not pick up differences in the waiting time distribution due the large uncertainty on this measure of the extreme. Moreover, the MWT does not consider the shape of the waiting time distributions,

which may differ in the thickness of the tails but have the same MWT. It is seen that *CVaR* picks up the change in the distribution of waiting times by using information from the whole tail rather than relying on the most extreme observation in each run.

5 Conclusions

The analysis of simulation studies needs reliable performance measures to answer the relevant research questions. In this article *CVaR* is suggested as a measure of the tail distribution of waiting times for a surgical unit with the objective of avoiding long waiting times. Our analysis shows that *CVaR* is a reliable measure that is specific to the tail. Moreover, *CVaR* can be seen as a compromise between the risk neutral average waiting time and the risk averse maximum waiting time. The results presented in this article show that using the maximum waiting time is a poor choice since it is highly variable and ignores changes in the shape of the waiting time distribution.

The average waiting time is not always representative for the waiting times, since such distributions often are skewed and long waiting times may potentially be more problematic from the points of view of patients and management. The *VaR* criterion is a measure of a quantile in the distribution but is indifferent to the tail distribution and does not quantify the tail distribution. In terms of quality management with patient satisfaction as outcome *CVaR* is highly relevant since it quantifies the problematic long waiting times. Moreover, the *CVaR* criteria

is more stable compared to *VaR* with respect to the chosen α-level since it is a sample average. It has nice properties as it is easy to compute and interpret and it is robust. *CVaR* of the waiting times may therefore be a relevant outcome in many quality improvement studies within health care with the objective of reducing the risk of long waiting times.

6 Author biographies

Christian Dehlendorff is a PhD-student in Informatics and Mathematical Modeling at the Technical University of Denmark. He has a M.Sc. in Engineering within data analysis and statistics. His research interests are within design of experiments and computer experiments.

Murat Kulahci is an Associate Professor in Informatics and Mathematical Modeling at the Technical University of Denmark. His research interests include design of experiments, statistical process control, and financial engineering. He is a member of the American Statistical Association, European Network of Business and Industrial Statistics (ENBIS), and the Institute of Operations Research and the Management Sciences.

Søren Merser is a surgeon (MD) at Clinic of Orthopedic Surgery at Frederiksberg Hospital, Denmark. He is a member of Danish Orthopedic Society and his primary research interest is on-line quality control in hospital units.

Klaus K. Andersen is an Associate Professor in Informatics and Mathematical Modeling at the Technical University of Denmark. He has a PhD in time series

analysis and his research interests are within design of experiments and statistical consulting.

References

[1] Alexander, S., Coleman, T. and Li, Y. (2006). Minimizing cvar and var for a portfolio of derivatives. *Journal of Banking and Finance*, 30(2), 583–605.

[2] Best, D. and Roberts, D. (1975). Algorithm as 89: The upper tail probabilities of spearman's rho. *Applied Statistics*, 24, 377–79.

[3] Bielen, F. and Demoulin, N. (2007). Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality*, 17(2), 174–193.

[4] Brailsford, S. C. (2007). Tutorial: Advances and challenges in healthcare simulation modelling. *Proceedings of the 2007 Winter Simulation Conference*, 1436–1448.

[5] Cayirli, T. and Veral, E. (2004). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4), 519–49.

[6] Conover, W. J. (1971). *Practical Nonparametric Statistics*. New York: John Wiley & Sons. Pages 295-301 (one-sample Kolmogorov test), 309-314 (two-sample Smirnov test).

[7] Denton, B. T., Rahman, A. S., Nelson, H. and Bailey, A. C. (2006). Simulation of a multiple operationg room surgical suite. *Proceedings of the 2006 Winter Simulation Conference*, 414–424.

[8] Ferrin, D. M. and McBroom, D. L. (2007). Maximizing hospital financial impact and emergence department throughput with simulation. *Proceedings of the 2007 Winter Simulation Conference*, 1566–1573.

[9] Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G. and Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health Medicine*, 25(4), 325–35.

[10] Garca-Gonzlez, J., Parrilla, E. and Mateo, A. (2007). Risk-averse profit-based optimal scheduling of a hydro-chain in the day-ahead electricity market. *European Journal of Operational Research*, 181(3), 1354–1369.

[11] Hollander, M. and Proschan, F. (1972). Testing whether new is better than used. *The Annals of Mathematical Statistics*, 78(4), 1136–1146.

[12] Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. New York: John Wiley & Sons. Pages 185-94.

[13] Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. New York: John Wiley & Sons. Pages 27-33 (one-sample), 68-75 (two-sample).

[14] Jun, J., Jacobson, S. and Swisher, J. (1999). Application of discrete-event simulation in health care clinics: a survey. *Journal of the Operational Research Society*, 50(2), 109–23.

[15] Kibzun, A. and Kuznetsov, E. (2003). Comparison of var and cvar criteria. *Automation and Remote Control*, 64(7), 153–164.

[16] Kibzun, A. I. and Kuznetsov, E. A. (2006). Analysis of criteria var and cvar. *Journal of Banking & Finance*, 30(2), 779–796.

[17] Krahl, D. (2002). The extend simulation environment. *Proceedings of the 2002 Winter Simulation Conference*, 205–213.

[18] Miller, M., Ferrin, D., Ashby, M., Flynn, T. and Shahi, N. (2007). Merging six emergency departments into one: A simulation approach. *Proceedings of the 2007 Winter Simulation Conference*, 1574–1578.

[19] Nakayama, M. K. (2006). Output analysis for simulations. *Proceedings of the 2006 Winter Simulation Conference*, 36–46.

[20] R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL `http://www.R-project.org`

[21] Rockafellar, R. T. and Uryasev, S. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26, 1443–1471.

[22] Sargent, R. G. (1998). Verification and validation of simulation models. *Proceedings of the 1998 Winter Simulation Conference*, 121–130.

[23] Sarkar, D. (2009). *lattice: Lattice Graphics*. R package version 0.17-22. URL `http://CRAN.R-project.org/package=lattice`

[24] Silber, J. H., Rosenbaum, P. R., Koziol, L. F., Sutaria, N., Marsh, R. R. and Even-Shoshan, O. (1999). Quality and outcomes of care - conditional length of stay. *Health Services Research*, 34(12), 349–363.

[25] Silverman, B. W. (1986). *Density Estimation*. Chapman and Hall. Page 48.

[26] Tang, L., Lu, Y. and Chew, E. (1999). Mean residual life of lifetime distributions. *IEEE Transactions on Reliability*, 48(1), 73–78.

[27] Webby, R., Adamson, P., Boland, J., Howlett, P., Metcalfe, A. and Piantadosi, J. (2007). The mekong-applications of value at risk (var) and conditional value at risk (cvar) simulation to the benefits, costs and consequences of water resources development in a large river basin. *Ecological Modelling*, 201(1), 89–96.

Paper B

Designing Simulation Experiments with Controllable and Uncontrollable Factors

Invited conference paper published in Proceedings of Proceedings of the 2008 Winter Simulation Conference, S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, J. W. Fowler eds.

Proceedings of the 2008 Winter Simulation Conference
S. J. Mason, R. R. Hill, L. Moench, O. Rose, eds.

DESIGNING SIMULATION EXPERIMENTS WITH CONTROLLABLE AND UNCONTROLLABLE FACTORS

Christian Dehlendorff
Murat Kulahci
Klaus Kaae Andersen

Department of Informatics and Mathematical Modelling
Technical University of Denmark
Bygning 321, Richard Petersens Plads
Lyngby, DK-2800, DENMARK

ABSTRACT

In this study we propose a new method for designing computer experiments inspired by the split plot designs used in physical experimentation. The basic layout is that each set of controllable factor settings corresponds to a whole plot for which a number of subplots, each corresponding to one combination of settings of the uncontrollable factors, is employed. The caveat is a desire that the subplots within each whole plot cover the design space uniformly. A further desire is that in the combined design, where all experimental runs are considered at once, the uniformity of the design space coverage should be guaranteed. Our proposed method allows for a large number of uncontrollable and controllable settings to be run in a limited number of runs while uniformly covering the design space for the uncontrollable factors.

1 INTRODUCTION

With the current advances in computing technology, computer and simulation experiments are increasingly being used to study complex systems for which physical experimentation is usually not feasible. Our case study involves a discrete event simulation model of an orthopedic surgical unit. The discrete event simulation (DES) model describes the individual patient's progress through the system and has been developed in collaboration with medical staff at Gentofte University Hospital in Copenhagen. The unit undertakes both acute and elective surgery and performs more than 4,600 operative procedures a year. While the patients come from various wards throughout the hospital, the main sources of incoming patients are the four orthopedic wards or the emergency care unit.

The simulation model is implemented in Extend version 6 (Krahl 2002) on a Windows XP platform and controlled from a Microsoft Excel spreadsheet with a Visual Basic for application script. The model consists of 3 main modules: The wards and arrival, the operating facilities, and the recovery and discharge. Interaction with the surrounding hospital is for example modeled with simplified processes using the same resources as the processes in the surgical unit (occupying the resources) and with the patients entering and exiting the model. Operating rooms, recovery beds, wards and staff are included in the model. The average run time for simulating 6 months (with one week of warm-up) operations is around 7 minutes. Typical outcomes are waiting times, patient throughput and the amount of overtime.

The simulation model has two sources of noise coming from variations in the uncontrollable factors (a.k.a. environmental factors in physical experimentation) and from changes in the seed controlling the random number generation process embedded in the simulation model. The controllable factors are for example the number of operating rooms and the number of surgeons, whereas the uncontrollable factors may include for example the arrival rate of acute patients and the time required to clean the operating rooms.

In this type of application, several issues need to be considered. First, the controllable factors tend to be numerous and often discrete. Moreover a single experiment usually takes several minutes to run. Therefore a simple exhaustive method, where all possible combinations of the factor settings are considered, is often computationally infeasible due to the exponentially increasing number of factor combinations. Furthermore, the settings of the uncontrollable factors, e.g. the acute patient arrival rate or the duration of surgical procedures, are also of interest and must be determined as they may influence the outcome of the simulations and hence the robustness of the simulation analysis.

The paper is organized in the following manner: Section 2 introduces design of computer experiments and defines the performance measure for the designs. Section 3 describes the proposed design method and contrasts it with other methods. In section 4 opportunities for future research are presented. Finally the main conclusions are summarized in section 5.

2 DESIGN OF COMPUTER EXPERIMENTS

2.1 Literature Review

A general discussion on the issues regarding the design and analysis of computer experiments can be found in Sacks et al. (1989), Santner, Williams, and Notz (2003) and Fang, Li, and Sudjianto (2006). The outputs from the computer experiments are often considered to come from a deterministic computer code. In such experiments, the classical design of experiment methods such as replication is deemed to be redundant as replication of an experiment, for example, yields exactly the same result (see Santner, Williams, and Notz (2003) and Fang, Li, and Sudjianto (2006)).

Experiments based on a simulation model often involve some stochastic component; making the output also stochastic. Kleijnen (2008) discusses the design and analysis of simulation experiments which typically have some sort of noise in the output. Therefore these experiments differ from the deterministic computer experiments. Furthermore, a typical simulation application will have both controllable and uncontrollable (environmental) factors, which should be handled differently. In these applications the aim is to manipulate the controllable factors so that the system is insensitive (robust) to changes in the uncontrollable factors. As described by Kleijnen (2008) and Sanchez (2000) the solution's robustness needs to be considered in order to obtain applicable solutions in systems with uncontrollable factors. That is, a good solution needs to perform well over the entire range of uncontrollable factors.

The original concept of robustness in physical systems is often attributed to Taguchi (1987). Taguchi's methods involve an inner array for the controllable factors and an outer array for the uncontrollable factors. In simulation studies, Kleijnen (2008) suggests using a crossed design, e.g. combining a central composite design (CCD) for the controllable factors and a Latin Hypercube Design (LHD) for the uncontrollable factors. In a crossed design the same set of subplots is used for each whole plot. However, as we will

show in this study, this may not be the most efficient way of running such experiments.

2.2 Simulation Model

Our basis is a discrete event simulation model generating output, $y = f(\mathbf{x_c}, \mathbf{x_e})$, for the settings for the s_c controllable factors, $\mathbf{x_c}$, and the settings for the s_e uncontrollable factors, $\mathbf{x_e}$. The objective is not only to select the settings, $\mathbf{x_c^*}$, such that the solution is robust to changes in the uncontrollable factor settings as described in p. 130-134 in Kleijnen (2008), but also to understand the variation coming from the changes in the uncontrollable factor settings.

Since little prior knowledge of both controllable and uncontrollable factors is available, we require that a good design is simultaneously uniform over the design space of the controllable and uncontrollable factors. In the following, we will assume that the uniform coverage of the design space of the controllable factors is already achieved and that we are only concerned with the uncontrollable factors.

Robustness studies in physical experimentation often involve split-plot designs (Montgomery 2005). We will therefore use similar terminology when robustness studies are performed using computer experiments. In classic split-plot designs, a set of experiments called whole-plots is designed so that for each whole-plot another set of experiments called subplots are run. In robustness studies, the settings of the controllable factors often constitute the whole-plots, whereas the settings of the uncontrollable factors constitute the subplots. In Table 1, a whole-plot corresponds to a row in which randomly selected combinations of settings for the uncontrollable factors are run. It should be noted that the randomization issue is irrelevant for computer experiments.

In the proposed method, each whole-plot corresponds to one combination of settings of the controllable factors (a row in Table 1), i.e. a total of n_c whole-plots are needed ($n_c = 5$ in Table 1). Each subplot (a column entry in any row in Table 1) corresponds to a combination of settings for the uncontrollable factors with a total of k subplots for each whole-plot. Thus the overall design consists of $N = n_c k$ runs. In a crossed design as proposed by Kleijnen (2008) these k subplots would be the same from one whole-plot to the next. Therefore there will only be a total of k combinations of settings for the uncontrollable factors. In our proposed methodology, different k combinations of settings for the uncontrollable factors will be used for each whole-plot. This is expected to give better overall coverage of the uncontrollable factor space compared to the crossed design. The challenge with the proposed method is to make the uncontrollable factor settings comparable from one whole-plot to the next.

Table 1: Uncontrollable factor design for five controllable settings and five environmental settings within each controllable setting

Controllable setting	Environmental setting				
	1	2	3	4	5
1	x_{e1}	x_{e2}	x_{e3}	x_{e4}	x_{e5}
2	x_{e6}	x_{e7}	x_{e8}	x_{e9}	x_{e10}
3	x_{e11}	x_{e12}	x_{e13}	x_{e14}	x_{e15}
4	x_{e16}	x_{e17}	x_{e18}	x_{e19}	x_{e20}
5	x_{e21}	x_{e22}	x_{e23}	x_{e24}	x_{e25}

2.3 Measure of Uniformity

In order to evaluate the designs presented in the following sections a measure of uniformity is needed. Fang, Li, and Sudjianto (2006) summarize a set of performance measures frequently used for measuring the uniformity of a design: the star discrepancy, centered discrepancy and the wrap-around discrepancy. The centered and the wrap-around discrepancy were proposed by Hickernell (1998b) and Hickernell (1998a), respectively. Both have desirable properties. They are easy to compute,

invariant to permutations of factors or runs and rotation of coordinates, and reliable measurements for the uniformity of projections. However the wrap-around discrepancy is said to be unanchored (i.e. it only involves the design points), while the centered discrepancy is not, since it involves the corners of the unit cube.

In this study only the wrap-around discrepancy is considered as the measure of uniformity with a low value corresponding to a high degree of uniformity. The measure is chosen since the literature generally suggests it as a good measure of uniformity (see for example Fang and Ma (2001); Fang, Lin, and Liu (2003); Fang, Li, and Sudjianto (2006)). The idea behind this measure is that for any two points from a uniform design, x_1 and x_2, spanning a hyper cube (potentially wrapping around the bounds of the unit cube); the hypercube should contain a fraction of the total number of points equal to the fraction of total volume covered by the cube. An analytic expression for the wrap-around discrepancy (WD(D)) is given by Fang and Ma (2001) as

$$(WD(D))^2 = -\left(\tfrac{4}{3}\right)^s + \tfrac{1}{n}\left(\tfrac{3}{2}\right)^s + \tfrac{2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\prod_{i=1}^{s} d_i(j,k) \quad (1)$$

with $d_i(j,k) = \frac{3}{2} - |x_{ki} - x_{ji}|(1 - |x_{ki} - x_{ji}|)$, n being the number of points, s the number of factors (the dimension), and x_{ki} the i'th coordinate of the k'th point.

There are various ways of constructing uniform designs. In this study the good lattice point method based on the power generator is used with the modification described in Fang, Li, and Sudjianto (2006). The design construction is based on a lattice $\{1,\ldots,n\}$ and a generator $h(k) = (1,k,k^2,\ldots,k^{s-1})(\text{mod } n)$, with k fulfilling that $k,k^2,\ldots,k^{s-1}(\text{mod } n)$ are distinct. $h(k)$ is chosen such that the resulting design consisting of the elements $u_{ij} = ih(k)_j(\text{mod } n)$ scaled down to $[0,1]^s$ has the lowest WD-value.

3 DESIGN ALGORITHM

A method for generating good designs for simulation models with both controllable and uncontrollable factors is presented in the following section. Here we assume that all factors have been scaled to $[0,1]$ and that the wrap-around discrepancy is the measure of uniformity. It is furthermore assumed that a design for the controllable factors is available. That is, we are primarily concerned with designing experiments for the uncontrollable factors. Two and three dimensional examples are used since they can be illustrated graphically. However, the method is general and results for 4 and 10 factors are also presented.

3.1 Bottom-up Approach

In section 2.2 the limitations of crossing a design for the controllable factors with a design for the uncontrollable factors were described. A better method in terms of covering the uncontrollable factor space compared to the crossed design is to generate different designs for the whole-plots, each with k different combinations of uncontrollable factor settings. This implies that n_c designs of size k should be constructed. For this method to succeed in the combined design, not only sets of k subplots for different whole-plots should be comparable, but also $n_c k$ subplots need to cover the design space for the uncontrollable factors uniformly. This can be achieved by dividing the design hyperspace for the uncontrollable factors into k sub-regions and sample n_c settings in each. As shown in Figure 1, this can be achieved fairly easily in two dimensions. However, in higher dimensions an efficient way of generating the sub-regions is required since the curse of dimensionality dictates that exponentially

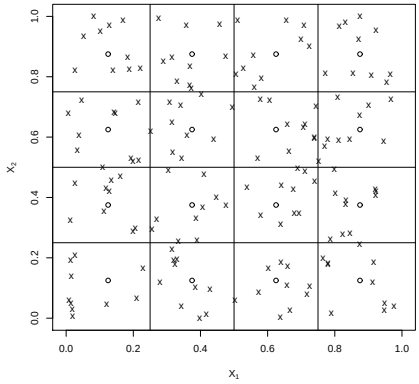


Figure 1: A total design of size 160 settings in 16 regions with 10 settings in each. Circles correspond to centers and crosses to sample settings.

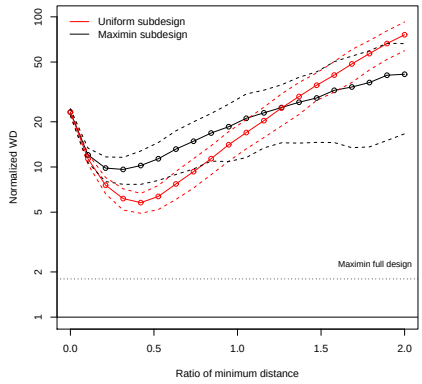


Figure 2: Average WD-value normalized using the WD-value obtained for a uniform design with 200 runs. Black curve with marks is for the maximum design and the red for the uniform design with dashed curves corresponding to approximate 95 % confidence intervals, the bottom black solid curve indicates a ratio of 1, i.e. no difference. The black dotted curve corresponds to a maxi-min distance. The overall design consists of 200 settings with the number of environmental settings being 40.

increasing numbers of runs have to be used in higher dimensions to obtain the same density of runs as in the lower dimensions.

If regular partitioning of the hypercube is possible, a design can be generated by randomly taking a run from each sub-region for each whole-plot. Figure 1 illustrates the approach in two dimensions with 16 subplots in each of the 10 whole plots. The design in Figure 1 has poor overall uniformity, which can also be seen from WD-values being 12 to 51 times higher compared to a uniform design of the same size.

A general method for generating the sub-regions is to generate a uniform design of size k and use these points as center points of k hypercubes or spheres that will constitute the sub-regions. The subplots are then generated within these sub-regions by either uniform designs or maxi-min distance designs for which the minimum distance of two runs in a sub-region is maximized. Figure 2 illustrates the performance of these methods for five controllable and 40 environmental settings for two environmental factors. The performance parameter in the figure is the WD-value for the combined environmental factor design, normalized by the WD-value of a uniform design of size 200. It can be seen that, compared to a uniform design generated directly for the same number of runs, both bottom-up methods are significantly worse. A maxi-min design generated directly is also seen to be better than the bottom-up generated designs. Figure 2 illustrates that using a bottom-up approach does not ensure an overall uniform design for the uncontrollable factors.

3.2 Top-down Approach

The second method we propose has more of a ”top-down” structure. First, we generate a

uniform design of size N which is equal to kn_c. This assures that the combined design is indeed uniform. But this does not solve the problem of assigning k settings to each of the n_c whole-plots such that in each whole-plot the subplots are uniformly spaced.

One approach to generate the designs is first to construct k sub-regions around k centers, where each region consists of n_c points. A method to obtain such a structure is to generate another uniform design of size k and use these points as starting center points, c, in an optimization algorithm that finds the optimal center points by minimizing

$$\sum_j \min_i ||x_j - c_i|| + k\sum_i (n_i - n_c)^2 \qquad (2)$$

In the above expression, n_i is the number of points having center i as the closest center. That is, the objective is to choose the centers, c^* such that they minimize the sum of the smallest differences between points and the centers, and the deviations from the required size of the region. This should ensure reasonably good separation of the points.

Based on the optimal centers, c^*, the N points need to be assigned to a center such that all points are assigned and all centers have exactly n_c points. This can be done in various ways, for example by assigning the point with the smallest distance to its nearest center, or by assigning the point with the largest second-shortest distance to its nearest center, or by simply considering the points' membership to each center based on euclidean distances.

A result of assigning 400 points to 10 groups of 40 points each is shown on the left of Figure 3, where it can be seen that the resulting groups are not well defined. Applying an exchange-algorithm on the assignment significantly improves the assignment as seen on the right of Figure 3. The total distances of the points to their center are reduced by 5 % by swapping less than 20 points and the points are grouped in well-defined clusters. An example in three dimensions is shown in Figure 4. The grouping in Figure 4 is generated by applying the exchange algorithm to a completely random assignment leading to a 49 % improvement in the distance of the points to the centers by more than 200 swaps.

3.2.1Generating Whole Plots

After grouping the subplots in k groups, we generate the whole-plots. Each whole-plot is assigned to one setting from each of the k groups so that all settings are assigned. One method is to assign the settings such that the maximum WD-value of the sub-designs is minimized, which can be obtained by repeatedly assigning the settings randomly to the whole-plots until a certain degree of uniformity is obtained.

Another method is to move the small uniform design of size k so that the point closest to the origin in the small design is placed at the points in the group closest to the origin and then assign points based on the smallest distance. The advantage of this approach compared to random assignment is that the whole-plot approximately mimics the uniform design structure.

For the designs considered in Figure 3 and 4 the performance of each whole-plot is compared to a uniform design generated directly in Table 2. The table shows that the overall uniformity of the combined design cannot be fulfilled without getting sub-designs that are not completely uniform. The designs with lowest maximum relative WD-value all have WD-values below 3.7 times and the highest minimum WD-values are less than twice the reference designs.

It can be seen from Table 2 that the results are consistent for up to 10 factors. The mean and the smallest maximum WD-value are all decreasing, whereas the remaining values are inconclusive with respect to the number of factors. It can also be seen from Table 2 that a design, which ensures relative WD-values for

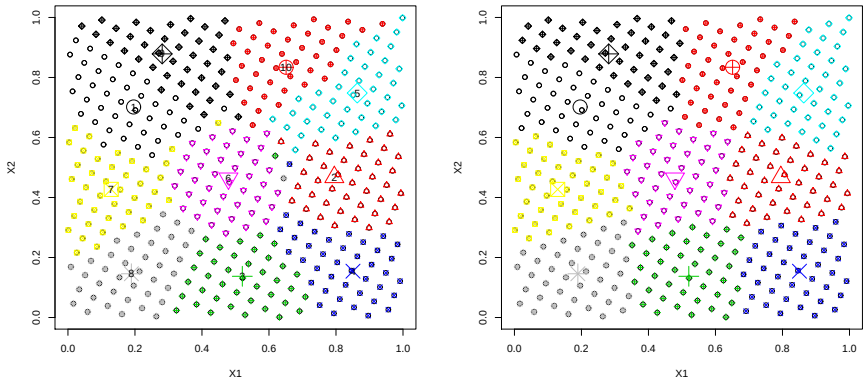


Figure 3: Left: The optimal assignment corresponding to a membership assignment. Right: The assignment after swapping in the optimal design.

all whole-plots between 2 (Max min) and 3.7 (Min max) can be achieved for up to 10 factors. The results seem to be independent of the number of settings but with 10 factors generally giving significantly lower values. This may be caused by the sparsity of the settings in the 10 dimensional design space.

4 DISCUSSION

This study was originated from application of discrete event simulation and computer experimentation at a hospital unit. In health- care applications in general, it is desireable that the final solutions are robust to changes in the uncontrollable factors. In the proposed design a large set of combinations of the uncontrollable factor settings is achieved using only a limited number of runs in each whole-plot. This is due to the fact that in each whole-plot a different set of subplots is used. When considered together, however, the subplots in the combined design show a uniform coverage of the design space.

Based on the proposed design, a meta-model of the following form

$$y(x_e, x_c) = f_1(x_c) + f_2(x_e) + f_{12}(x_c, x_e) + e \quad (3)$$

could be considered with $f_1(x_c)$ being a function describing the fixed effects related to the controllable setting, $f_2(x_e)$ and $f_{12}(x_c, x_e)$ being random effects describing the variations on the mean effect and the effect of the uncontrollable factor variations on the fixed effects.

By ensuring the overall uniformity of the uncontrollable factor settings, the functions $f_2(x_e)$ and $f_{12}(x_c, x_e)$ can be estimated over the whole region. The functions $f_2(x_e)$ and $f_{12}(x_c, x_e)$ describe the impacts of the variations in the uncontrollable factors. These can be used for quality improvement purposes if the variation in some of the uncontrollable factors is somehow possible to reduce. Moreover, $f_{12}(x_c, x_e)$ is of interest in robustness studies since the interaction between controllable and uncontrollable factors is the key to reducing the impact from changes in the uncontrollable factors.

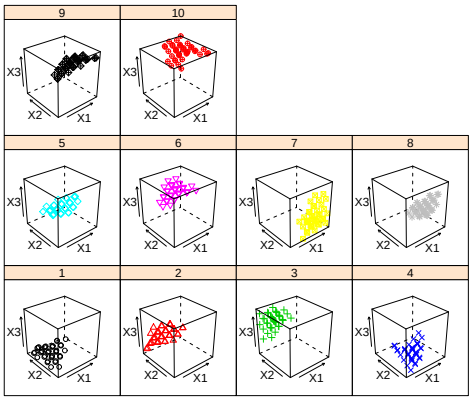


Figure 4: 400 settings assigned to 10 groups in 3 dimensions. Each panel corresponds to one group.

Table 2: Summary for relative WD-values for 2, 3 and 4 dimensional examples with 40 controllable factors, each with 10 environmental settings (400) or 20 controllable factors, each with 10 environmental settings (200). The performance is summarized by minimum (Min), mean (Mean) and maximum (Max) relative WD-value and by the highest minimum (Max min) and lowest maximum (Min max). The values are relative to the WD-value for a uniform design of the same size as the whole-plots

Factors	Min	Max min	Mean	Min max	Max
2 (400)	1.15	1.99	2.78	3.67	8.39
3 (400)	1.19	1.93	2.70	3.47	7.21
4 (400)	1.25	1.94	2.56	3.20	7.28
10 (400)	1.32	1.60	1.76	2.00	2.38
2 (200)	1.14	2.17	2.69	2.94	7.20
3 (200)	1.17	2.21	2.68	2.94	6.98
4 (200)	1.22	2.22	2.50	2.54	5.65
10 (200)	1.29	1.63	1.73	1.78	2.45

5 CONCLUSION

In this study, a methodology to design uniformly distributed experiments for simulation experimentation in the presence of both controllable and uncontrollable factors is introduced. The method ensures that the subplots in the combined design for the uncontrollable factors are uniform while keeping an acceptable level of uniformity of the subplots within each whole-plot. Complete uniformity compared to uniform design of the size equal to the total number of subplots could not, however, be achieved.

The proposed methodology is primarily based on Euclidian distances. Therefore the method can be used in designs with many uncontrollable/environmental factors. Our results show that a uniformity measure of the individual whole-plots can be minimized to within two to four times the value of an overall uniform design. Furthermore, it was shown that the method was applicable to designs with 2 to 10 uncontrollable factors. Since the methodology is based on distances, increasing the number of factors may be possible, although sparsity of the experiments in the design space may become an issue.

The proposed design contains as many uncontrollable factor settings as the number of runs (N), which in contrast to a crossed design of the same size has $k = N/n_c$ unique uncontrollable factor settings. This implies that the simulation time for a crossed design with the same number of unique uncontrollable factor settings becomes n_c times longer. For a fixed experimental design size, the proposed design optimally covers the uncontrollable factor space in terms of overall uniformity. In the modeling and analysis of the simulation output, the uniformity provides good coverage for the uncontrollable factor effects.

AUTHOR BIOGRAPHIES

CHRISTIAN DEHLENDORFF is a Ph.D. student at the Department of Informatics and Mathematical Modelling, Technical University of Denmark. His email and web addresses are <cd@imm.dtu.dk> and <http://www.imm.dtu.dk/~cd>.

MURAT KULAHCI is an Associate Professor at the Department of Informatics and Mathematical Modelling, Technical University of Denmark. His email address is <mk@imm.dtu.dk>.

KLAUS KAAE ANDERSEN is an Associate Professor at the Department of Informatics and Mathematical Modelling, Technical University of Denmark. His email address is <kka@imm.dtu.dk>.

REFERENCES

Fang, K.-T., R. Li, and A. Sudjianto. 2006. *Design and modeling for computer experiments*. Chapman & Hall/CRC.

Fang, K.-T., D. K. J. Lin, and M.-Q. Liu. 2003. Optimal mixed-level supersaturated design. *Metrika* 58 (3): 279–291.

Fang, K.-T., and C.-X. Ma. 2001. Wrap-around l2-discrepancy of random sampling, latin hypercube and uniform designs. *Journal of Complexity* 17 (4): 608–624.

Hickernell, F. 1998a. *Random and quasi-random point sets*, Chapter Lattice rules: How well do they measure up?, 106–166. Springer-Verlag, New York.

Hickernell, F. J. 1998b. A generalized discrepancy and quadrature error bound. *Mathematics of Computation* 67 (221): 299–322.

Kleijnen, J. P. 2008. *Design and analysis of simulation experiments*. Springer.

Krahl, D. 2002. The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*, 205–213.

Montgomery, D. C. 2005. *Design and analysis of experiments*. 6th ed. John Wiley and Sons, Inc.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Science* 4 (4): 409–423.

Sanchez, S. M. 2000. Robust design: Seeking the best of all possible worlds. In *Proceedings of the 2000 Winter Simulation Conference*, 69–76.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The design and analysis of computer experiments*. Springer.

Taguchi, G. 1987. *System of experimental design, volumes 1 and 2*. UNIPUB/Krauss International, White Plains, New York.

Paper C

Designing simulation experiments with controllable and uncontrollable factors for applications in health care

---

Accepted for publication in Journal of Royal Statistical Society series C 2011

Designing simulation experiments with controllable and uncontrollable factors for applications in health care

Christian Dehlendorff
Murat Kulahci
Klaus Kaae Andersen

DTU Informatics, Technical University of Denmark
DK-2800 Lyngby
Denmark

Summary.
We propose a new methodology for designing computer experiments inspired by the split plot designs often used in physical experimentation. The methodology has been developed for a simulation model of a surgical unit in a Danish hospital. We classify the factors as controllable and uncontrollable based on their characteristics in the physical system. The experiments are designed so that for a given setting of the controllable factors, the various settings of the uncontrollable factors cover the design space uniformly. Moreover the methodology allows for overall uniform coverage in the combined design when all settings of the uncontrollable factors are considered at once.

Keywords: Computer Experiments, Design of Experiments, Discrete Event Simulation, Uniform design, Robustness

1. Introduction

With the current advances in computing technology, computer and simulation experiments are increasingly being used to study complex systems for which physical experimentation is usually not feasible. Our case study involves a discrete event simulation model of an orthopedic surgical unit at Gentofte University Hospital in Copenhagen. The discrete event simulation (DES) model describes the individual patient's progress through the system and has been developed in collaboration with medical staff at the hospital. The surgical unit undertakes both acute and elective surgery, and performs more than 4,600 operative procedures a year. Even though the patients come from several wards throughout the hospital, the main sources of incoming patients are four orthopedic wards and the emergency care unit. The patient's route through the unit is sketched in Figure 1.

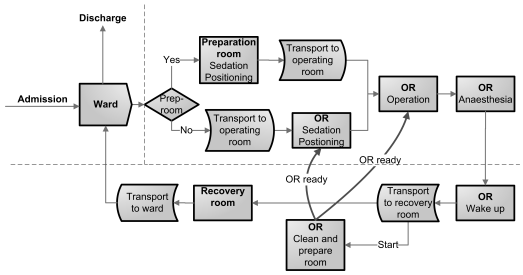


Fig. 1. Patient route through orthopedic surgical unit

The simulation model consists of three main modules: the wards (admission and discharge), the operating facilities, and the recovery. Interaction with the surrounding hospital is modeled as auxiliary processes using the same resources as the processes in the surgical unit (occupying the resources) and with the flow of patients between the unit and the rest of the hospital. Resources in the model are operating rooms, recovery beds and wards. The staff is also included in the model as a resource and controlled by resource pools. Outpatients treated in outpatient clinics are not considered in this model but the resources shared between outpatient clinics and the surgical unit are included.

The simulation model is implemented in Extend version 6 (Krahl, 2002) on a Windows XP platform and controlled from a Microsoft Excel spreadsheet with a Visual Basic for application script. The average run time for simulating six months' (with one week of warm-up) operation is approximately seven minutes excluding summarizing the run. Typical outcomes are waiting times, patient volume and amount of overtime. Waiting time is defined as the time a patient unnecessarily waits between procedures and it is closely related to patient satisfaction as described in Bielen and Demoulin (2007). As patient waiting time and patient satisfaction are the primary concerns, we restrict our focus to the patient waiting times; i.e., a single performance measure.

The simulation model has two sources of noise: external noise Ankenman et al. (2010) coming from variations in the uncontrollable factors (a.k.a. environmental factors in physical experimentation) and internal noise coming from changes in the seed controlling the random number generation process embedded in the simulation model. In addition, a set of controllable factors influence the system in a deterministic manner. The controllable factors are, for example, the number of recovery beds and the number of anesthesiologists, whereas the uncontrollable factors include the arrival rate of acute patients and the amount of time the recovery beds and anesthesiologists are being used by other processes.

In this type of application, several issues need to be considered. First, the controllable factors tend to be numerous and often discrete. Moreover a single experiment takes several minutes to run, and simple exhaustive methods, where all possible combinations of the factor settings are considered, are computationally unfeasible due to the exponentially increasing number of factor combinations. Thus, the selected factor combinations for experimentation must be chosen carefully. The second issue is that the settings of the uncontrollable factors are also of interest and must be analyzed, as their effect may influence the outcome of the simulations.

This paper is organized in the following manner: Section 2 introduces design of computer experiments and defines the performance measure for the designs. Section 3 describes the proposed design method and contrasts it with other methods. The design is illustrated by experimentation on the case study in section 4. In section 5 possible areas for future research are presented. Finally the main conclusions are summarized in section 6.

2. Design of computer experiments

A general discussion on the issues related to the design and analysis of computer experiments can be found in Sacks et al. (1989), Santner et al. (2003) and Fang et al. (2006). The main characteristic of computer experiments is that output is most often considered to come from a deterministic computer code. In such experiments, the classical design of experiment methods such as replication, randomization and blocking are deemed to be redundant (see Santner et al. (2003) and Fang et al. (2006)).

Experiments based on simulation models often involve some stochastic component; making the output also stochastic. Kleijnen (2008, 2009) discusses the design and analysis of simulation experiments which typically have some sort of noise in the output. Therefore these experiments differ from deterministic computer experiments. As in the case of physical experimentation, a typical simulation application will have both controllable and uncontrollable (environmental) factors. In these applications the aim is to manipulate the controllable factors so that the system is insensitive (robust) to changes in the uncontrollable factors. As described by Kleijnen (2008) and Sanchez (2000) the solution's robustness needs to be considered in order to obtain applicable solutions in systems with uncontrollable factors.

The original concept of robustness in physical systems is often attributed to Taguchi (1987). Taguchi's methods involve an inner array for the controllable factors and an outer array for the uncontrollable factors. In simulation studies, Kleijnen (2008, 2009) suggests using a crossed design, e.g., combining a central composite design (CCD) for the controllable factors and a Latin Hypercube Design (LHD) for the uncontrollable factors. In a crossed design the same set of uncontrollable factor settings is used for each controllable factor setting. However, as we will show in this study, it can be argued that this may not be the most efficient way of running such experiments.

2.1. Simulation model

We consider a discrete event simulation model generating output, $y = f(\mathbf{x_c}, \mathbf{x_u})$, for the settings for the s_c controllable factors given in $\mathbf{x_c}$ and the settings for the s_u uncontrollable factors given in $\mathbf{x_u}$. The objective is not only to select the settings, $\mathbf{x_c^*}$, such that the solution is robust to changes in the uncontrollable factor settings as described in Kleijnen (2008, p. 130-134), but also to provide insight into how the variation coming from changes in the uncontrollable factor settings causes variation in the output.

In the following, we will assume that an experimental plan for the controllable factors is already available (for example, a factorial design) so that we are only concerned with choosing the uncontrollable factor settings. Because little prior knowledge of the effects of these factors is usually available, we require that the factor space for the uncontrollable factors is uniformly covered for each controllable factor setting (the sub-designs) as well as in the combined design for which all uncontrollable factor settings are considered at once. Overall uniformity is important for the robustness of the analysis (Fang et al., 2006) and the uniformity of the sub-designs is required in order to achieve similar environmental variations for all combinations of the

controllable factor settings. Another objective of the experiment plan could be to generate informative data for building computationally less expensive surrogates for the simulation models.

Robustness studies in physical experimentation often involve split plot designs (Montgomery, 2009). We will apply a similar terminology when robustness studies are performed using computer experiments. In classical split plot designs, a set of experiments called whole plots is designed so that for each whole plot another set of experiments called subplots is run. In robustness studies, the settings of the controllable factors often constitute the whole plots, whereas the settings of the uncontrollable factors constitute the subplots. In Table 1, a whole plot corresponds to a row in which randomly selected combinations of settings for the uncontrollable factors are run.

In physical experimentation, the whole plots and subplots are randomized separately; that is, for each randomly selected whole plot, corresponding subplots are run in a random order. The separate randomization of whole plots and subplots is typically applied when the whole plot factors are hard to change; i.e., keeping them at a fixed level while varying the subplot factors makes the experiment less time consuming and/or expensive. Our design is not a split plot experiment, but it has some similarities in the structure. For computer and simulation experiments the randomization is not an issue, since everything is controlled. To ease the notation in the rest of the paper, we will use a whole plot for a setting of the controllable factors and a subplot for a setting of the uncontrollable factors.

In the proposed method, each whole plot corresponds to one combination of settings of the controllable factors (a row in Table 1); i.e., a total of n_c whole plots are needed ($n_c = 5$ in Table 1). Each subplot (a column entry in any row of Table 1) corresponds to a combination of settings for the uncontrollable factors with a total of k subplots for each whole plot. Thus, the unreplicated overall design consists of $N = n_c k$ runs. In a crossed design as proposed by Kleijnen (2008) these k subplots would be the same from one whole plot to the next. Therefore there will only be a total of k unique combinations of settings for the uncontrollable factors in a crossed design. In our proposed methodology, different k combinations of settings for the uncontrollable factors will be used for each whole plot. This is expected to give better overall uniform coverage of the uncontrollable factor space compared to the crossed design, which is thought to be of increasing importance as the number of uncontrollable factors increases. One of the greatest challenges with the proposed method is to make the variations in the uncontrollable factor settings comparable from one whole plot to the next.

2.2. Measure of uniformity

In order to evaluate the proposed designs, a measure of uniformity is needed. Fang et al. (2006) summarize a set of performance measures frequently used for measuring the uniformity of a design: the star discrepancy, the centered discrepancy and the wrap-around discrepancy. The centered and the wrap-around discrepancies were proposed by Hickernell (1998b) and Hickernell (1998a), respectively, and both have desirable properties. They are

Table 1. Uncontrollable factor design for five controllable settings and five environmental settings within each controllable setting

Controllable setting	Environmental setting				
	1	2	3	4	5
1	x_{e1}	x_{e2}	x_{e3}	x_{e4}	x_{e5}
2	x_{e6}	x_{e7}	x_{e8}	x_{e9}	x_{e10}
3	x_{e11}	x_{e12}	x_{e13}	x_{e14}	x_{e15}
4	x_{e16}	x_{e17}	x_{e18}	x_{e19}	x_{e20}
5	x_{e21}	x_{e22}	x_{e23}	x_{e24}	x_{e25}

easy to compute, invariant to permutations of factors, runs and rotation of coordinates, geometrically interpretable, and reliable measurements for the uniformity of projections. However, the wrap-around discrepancy is said to be unanchored (i.e. it only involves the design points), while the centered discrepancy also involves the corners of the unit cube. The computational costs of the star discrepancy make this criterion unsuitable as a uniformity measure (Fang et al., 2006).

Fang et al. (2006) do not give any recommendations for whether to choose the centered discrepency or the wrap-around discrepency. In this study we consider the wrap-around discrepancy, since it has the same desirable properties as the centered discrepancy, but involves the design points only and not the corner points as mentioned above. However, the method is not limited to this particular uniformity measure. The idea behind this measure is that for any two points from a uniform design, x_1 and x_2, spanning a hypercube (potentially wrapping around the bounds of the unit cube), the hypercube should contain a fraction of the total number of points equal to the fraction of total volume covered by the cube. An analytic expression for the wrap-around discrepancy for an experimental plan D is given by Fang and Ma (2001) as

$$(\mathrm{WD}(D))^2 = -\left(\tfrac{4}{3}\right)^s + \tfrac{1}{n}\left(\tfrac{3}{2}\right)^s + \tfrac{2}{n^2}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\prod_{i=1}^{s} d_i(j,k) \qquad (1)$$

with $d_i(j,k) = \frac{3}{2} - |x_{ki} - x_{ji}|(1 - |x_{ki} - x_{ji}|)$, n being the number of points, s the number of factors (the dimension), and x_{ki} the i'th coordinate of the k'th point. It is required that $x_{ki} \in [0,1]$, which shows that $d_i(j,k)$ is maximal with a distance of 0 or 1 between x_{ki} and x_{ji} and minimal with a distance of 0.5. A low WD value corresponds to a high degree of uniformity. For more details about the properties of WD, see for example Fang and Ma (2001), Fang et al. (2003) and Fang et al. (2006).

There are various ways of constructing uniform designs. In this study the good lattice point method based on the power generator is used with the modification described in Fang et al. (2006). The design construction is based on a lattice $\{1,\ldots,n\}$ and a generator $\mathbf{h}(k) = (1, k, k^2, \ldots, k^{s-1})(\mathrm{mod}\ n)$, with k fulfilling that $k, k^2, \ldots, k^{s-1}(\mathrm{mod}\ n)$ are distinct. The generator $\mathbf{h}(k)$ is chosen such that the resulting design consisting of the elements $u_{ij} = ih(k)_j(\mathrm{mod}\ n)$ scaled down to $[0,1]^s$ has the lowest WD value.

3. Design algorithm

In this study we will assume that all factors have been scaled to be in the interval $[0, 1]$ and that a design for the controllable factors is available; that is, we are primarily concerned with designing experiments for the uncontrollable factors. A two dimensional example is used as the primary example, since it can be easily visualized. However, the method is general and results for 3 to 19 factors are also presented.

In section 2.1 the limitations of crossing a design for the controllable factors with a design for the uncontrollable factors were discussed. A better method in terms of improving the coverage of the uncontrollable factor space compared to the crossed design is to generate different designs for the whole plots, each with k different combinations of uncontrollable factor settings. This implies that n_c designs of size k should be constructed. For this method to succeed in the combined design, not only should sets of k subplots for different whole plots be comparable, but also when the combined design is considered as a whole, the $n_c k$ subplots should cover the design space for the uncontrollable factors uniformly.

In Dehlendorff et al. (2008) we analyzed a "bottom-up" approach in which the overall design is constructed by splitting the hypercube spanning the uncontrollable factor space into k sub-regions. These k sub-regions are constructed so that each contains n_c points. We then select one point from each sub-region to form a set of k points and assign those to a single whole plot. The main problem with this construction method is that the overall uniformity of the combined design cannot be guaranteed. For a two dimensional example this yields WD values at least five-times higher than a uniform design generated directly for the entire uncontrollable factor space.

3.1. Top-down approach

The method we propose here has more of a "top-down" structure. First, we generate a uniform design of size $N = kn_c$ in the uncontrollable factor space. This assures that the combined design will indeed be uniform. But this does not solve the problem of assigning k settings of the uncontrollable factors to each of the n_c whole plots such that in each whole plot the subplots are uniformly spaced.

One approach to generate various k settings is first to construct k sub-regions around k centers, where each region consists of n_c points. A method to obtain such a structure is to generate another uniform design of size k in the hyperspace for the uncontrollable factors and use these points as starting center points, $C = \{\mathbf{c_1}, \ldots, \mathbf{c_k}\}$, in an optimization algorithm that finds the optimal center points as

$$C^* = \{\mathbf{c_1}, \ldots, \mathbf{c_k}\}^* = \underset{\{\mathbf{c_1},\ldots,\mathbf{c_k}\}}{\arg\min} \sum_{j=1}^{N} \min_i ||\mathbf{x_j} - \mathbf{c_i}|| + k \sum_{i=1}^{k} (n_i - n_c)^2 \qquad (2)$$

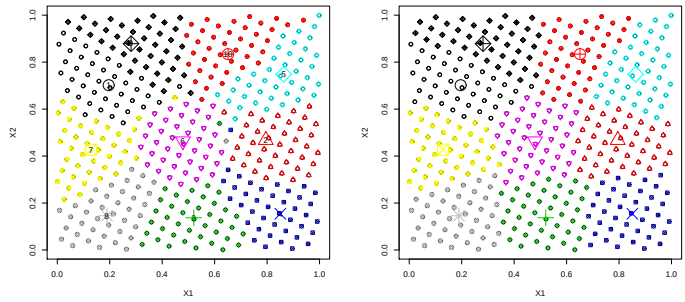
where n_i is the number of points having center $\mathbf{c_i}$ as the closest center; that is, the objective is to choose the centers, C, such that they minimize the sum of the smallest differences between points and their respective centers, and the

deviations from the required size of the region. This should ensure reasonably good separation of the points.

On the basis of the optimal centers, C^*, the N points need to be assigned to a center such that all points are assigned and all centers have exactly n_c points. This can be done in various ways, for example by simply considering the points' membership to each center based on Euclidean distances and then assigning them to their closest center (if the center has fewer than n_c points assigned already). The results of this initial grouping may be that some groups are not well defined; i.e., have points separated from the core of the group. In order to obtain well defined regions some sort of exchange algorithm may be needed after the initial grouping. One way to obtain more well defined regions is to swap the centers of two points as long as the total distance between points and their center becomes smaller. For example, we would exchange the centers for the points x_i and x_j if

$$\Delta_{ij} = [d(x_i, c(x_i)) + d(x_j, c(x_j))] - [d(x_i, c(x_j)) + d(x_j, c(x_i))] > 0 \qquad (3)$$

where $c(x_i)$ is the location of x_i's center and $d()$ measures the Euclidean distance. The implemented algorithm chooses the pair of points giving the highest reduction in each iteration and terminates when no further reduction is possible; i.e., $\Delta_{ij} \leq 0 \quad \forall i, j$.



(a) (b)

Fig. 2. (a): The optimal assignment corresponding to a membership assignment. (b): The assignment after applying an exchange algorithm to the optimal design.

The resulting scheme of assigning 400 points to 10 groups of 40 points each is shown in Figure 2(a), where it can be seen that the resulting groups are not well defined, e.g., group 3 in Figure 2(a) has a single point placed between groups 2, 5, 6 and 10. Applying the exchange algorithm on the initial grouping improves the tightness of the groups, as seen in Figure 2(b). The total distances of the points to their center are reduced by 5 % by swapping less than 20 points and the points are now grouped in well defined clusters. An example in three dimensions leads to a 49 % improvement in the distance of the points to the centers by approximately 200 swaps after a random initial assignment.

3.2. Generating whole plots

After grouping the subplots into k groups, of course the next question is about effectively assigning subplots for each whole plot. In the previous section we showed how to efficiently group the subplots in k groups of n_c points each. For a given group of n_c points, we assign each subplot to a whole plot so that all n_c subplots of a group are assigned to n_c distinct whole plots. In the assignment of the subplots we want the resulting n_c designs (sub-designs), corresponding to the n_c whole plots, to be as uniform as possible. One way is to choose the assignment minimizing the maximal (min-max) WD value of the sub-designs, and this reduces the risk of getting a sub-design with a low degree of uniformity.

Assignment of the points can be done by repeatedly assigning the subplots within each region randomly to the n_c whole plots and then choosing the assignment giving the lowest min-max value. However, this strategy becomes computationally intensive for a large number of subplots. Another method is to mimic the structure of the uniform design for the k centers used as starting points for the minimization in equation (2). This can be achieved by, for each of the n_c whole plots, superimposing the same uniform design of size k as used for construction of the center points on the combined design of size N such that the point closest to the origin in the design of size k matches one of the n_c subplots (the anchoring point) in the region closest to the origin. Having superimposed the design of size k, the i'th whole plot is generated by assigning, in each of the k regions, the subplot (which is not already assigned) closest to the superimposed design, such that the i'th whole plot is assigned exactly one subplot from each region. This can be repeated by choosing different sequences of subplots as anchoring points until the best assignment is chosen. A top-down design with n_c whole plots with s_c factors and k subplots with s_u factors is denoted $\mathrm{TD}(n_c, s_c, k, s_u)$.

We summarize the procedure of constructing the top-down design in the following steps

(a) Generate uniform design (U_b) with $N = n_c k$ points and s_u factors
(b) Split U_b into k sub-regions with n_c points each as follows

  (i) generate uniform design (U_s) with k points and s_u factors
  (ii) use U_s as starting points for optimizing equation (2) for C^*
  (iii) assign n_c points to each center by considering the Euclidean distances
  (iv) exchange centers as long as equation (3) is valid for a pair of points

(c) Assign k points to n_c whole plots as follows

  (i) find sub-region closest to the origin (i)
  (ii) find point in U_s closest to the origin (j)
  (iii) set current whole plot number to 1
  (iv) superimpose U_s on U_b such that the j'th point in U_s is placed in a random point not already assigned in the i'th sub-region of U_b
  (v) in each sub-region assign the point closest to U_s (if not already assigned) to the current whole plot

Table 2. Whole plot performance for different numbers of uncontrollable factors (s_u) and different numbers of overall number of subplots (N). The whole plot size is kept fixed at $k = 10$ corresponding $n_c = 20$ and $n_c = 40$ for $N = 200$ and $N = 400$, respectively. The performance for the n_c whole plot is summarized in the max-min corresponding to the highest minimum relative WD value and the min-max corresponding to the smallest maximum .

s_u	N	max-min	min-max	N	max-min	min-max
2	200	1.95	2.84	400	1.65	3.08
3	200	2.29	4.21	400	2.01	5.24
4	200	2.37	3.99	400	2.10	4.81
5	200	2.75	3.43	400	2.72	3.94
6	200	2.67	3.14	400	2.66	3.82
7	200	2.32	2.82	400	2.39	3.30
8	200	2.21	2.62	400	2.26	2.92
9	200	2.08	2.39	400	2.01	2.69
10	200	1.82	2.08	400	1.97	2.51
11	200	1.67	1.83	400	1.73	2.09
12	200	1.58	1.71	400	1.58	1.92
13	200	1.42	1.54	400	1.46	1.69
14	200	1.41	1.53	400	1.41	1.67
15	200	1.35	1.44	400	1.37	1.54
16	200	1.30	1.38	400	1.29	1.51
17	200	1.27	1.34	400	1.27	1.41
18	200	1.22	1.27	400	1.24	1.35
19	200	1.20	1.24	400	1.21	1.32

(vi) if current whole plot number is n_c then stop, otherwise increment current whole plot number by 1 and go to step c(iv)

(d) repeat step c and keep best assignment according to the min-max value, $\mathrm{TD}(n_c, s_c, k, s_u)$

For each combination of s_u and N, the sub-designs are summarized in Table 2 in terms of the maximal minimum (max-min) relative WD value (relative to a uniform design of size k generated directly for the same region) of the k sub-designs and the minimal maximum relative WD value (min-max). This implies that a design with all sub-design WD-values lying between the max-min and min-max can be constructed. Table 2 shows that the overall uniformity of the combined design cannot be fulfilled without getting sub-designs that are not completely uniform. The designs with lowest maximum relative WD value all have WD values less than 5.3 times the reference designs and the highest minimum WD values are less than three times the WD values of the reference designs. For the design considered in Figure 2(b) the performance of each whole plot is compared to a uniform design generated directly in Table 2 for $s_u = 2$ and $N = 400$, and shows that the uniformity of the whole plots is between 1.65 and 3.08 higher than of a comparable uniform design generated directly.

It can be seen from Table 2 that the results are consistent for up to 19 factors. The max-min value is highest for 5 factors, whereas the min-max value is highest for 3 factors. It can also be seen from Table 2 that a design that ensures relative WD values for all whole plots between 2.8 (max-min) and 5.3 (min-max) can be achieved for up to 19 factors. The values for max-min and min-max tend to go down with increasing s_u. This could be caused by the increasing sparsity in higher dimensions.

Table 3. Controllable factors for simulation experiment. Current corresponds to the current setting at the surgical unit

Factor	Low	High	Current
Anesthesiologists (A)	2	3	2
Porters (B)	3	4	3
Recovery beds (C)	6	8	6
Operating days (D)	5	4	5

4. Case study

To illustrate the advantages of using the top-down design, two different experiments with the simulation model are studied. The first experimental plan is a crossed design between n_c controllable factor settings and k uncontrollable factor settings. The results from this design are compared to the results from a top-down design of the same size.

We consider four controllable factors with two levels, each as shown in Table 3. The variable *Operating days* is constructed such that the number of weekly hours remains the same irrespective of the number of *Operating days*. The remaining three factors are related to the staffing during regular hours. Moreover, the levels are organized such that the current setting is the reference (low level) for all factors, which for *Operating days* implies that five days is the low level and four days the high level. For the controllable part of the design a 2^4 factorial design is employed (Montgomery, 2009); i.e., $n_c = 16$.

Because an important goal is to analyze the system performance under challenging settings of the uncontrollable factors, they are varied around their current estimated settings from a 20 % better scenario to a 50 % worse for each. This implies that the majority of the scenarios will have more challenging operating conditions compared to the current estimated settings. We select $k = 10$ uncontrollable factor settings for each controllable factor setting.

For the crossed design, a uniform design with $k = 10$ runs and eight uncontrollable factors is constructed and crossed with the 2^4 factorial experiment for the controllable factors. Moreover, a TD$(16, 4, 10, 8)$ is also constructed; i.e., a top-down design of the same size as the crossed design. This gives a total of two experimental plans, each with 160 ($= 16 \times 10$) runs, together requiring around 40 hours of simulation time.

Even though the uncontrollable factors used in our example come from a thorough study of the real system, we suspect (and to some extent expect) that the list is incomplete. For the effects of "unknown" factors that may have an effect, albeit small, on the response and hence creating additional noise, we choose to use random seed in our simulation model causing our simulations to become stochastic rather than deterministic. Hence a robust setting should not only be robust against the uncontrollable factors, it should also be robust against the intrinsic uncertainty introduced by the queues and procedures. The commonly used variance reduction technique of using common random numbers was tested, but gave similar results and did not give a clear-cut reduction in the variance of the estimates in section 4.2. Moreover, using different seeds implies that the observations can be assumed to be independent and this means that standard techniques can be applied.

As the response, we primarily focus on long patient waiting times measured by the average of the $\alpha = 5$ % longest waiting times. This corresponds to the conditional value at risk (CVaR), which is frequently used in finance (see e.g., Kibzun and Kuznetsov, 2003; Alexander et al., 2006). Dehlendorff et al. (2010) compared CVaR to other measures in the literature and found that CVaR was a reliable measure of the tail distribution of waiting times. The main advantage of using CVaR compared to, for example, the average or the maximum waiting time is that it is related to the distribution of the tail, whereas the average waiting time is based on the whole distribution and the maximum waiting time is a measure of an extreme. The two α-extremes 0 % and 100 % for CVaR correspond to the maximum and the average waiting time respectively, and CVaR forms a compromise between the two. In finance the average and the maximum waiting time correspond to risk-neutral and risk-averse strategies, respectively.

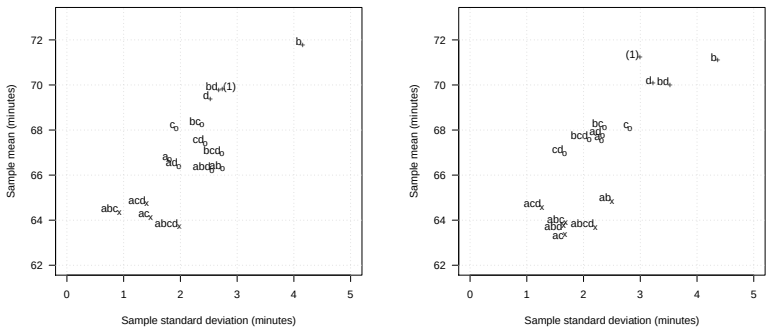
4.1. Taguchi approach

In Figure 3 the standard deviations and sample averages for each controllable factor setting (whole plot) are plotted for each of the designs. The results are similar with some minor differences, however as shown in Figure 4 and in the analysis based on equation (4) the difference is in the estimation of the uncontrollable factors. It can be seen that the crossed design (Figure 3(a)) has four settings in the lower left corner (marked with x) and the top-down design (Figure 3(b)) has the same four plus an additional two settings. These settings give both low and reliable waiting times. It can be seen that factor A is at its high level, indicated by a, in all settings having both low average and standard deviation, i.e. the anesthesiologist resource is potentially an important factor in obtaining consistently low waiting times. Likewise the four settings in the upper right corners of Figure 3(a) and 3(b) have factor A at its low level.

Taguchi (1987) uses the signal-to-noise ratio as the robustness measure in systems with controllable and uncontrollable factors. It is given as $SN = 20\log(\bar{y}/\bar{s})$, where $\bar{y}$ is the sample average for a given setting of the controllable factors and $\bar{s}$ the sample standard deviation. Taguchi proposes the signal-to-noise ratio as a trade-off between high mean and low uncertainty to quantify the robustness of a system. Using SN on the sample averages and standard deviations in Figure 3 gives different optimal solutions for the two designs; i.e., the top-down design suggests that acd is the optimal setting, whereas the crossed design suggests that abc is the optimal setting. Bursztyn and Steinberg (2006) point out that using signal-to-noise is not an optimal way to assess the robustness of the system, instead they recommend that the noise factors are included in the analysis, and this is considered in the following. The main drawback of the signal-to-noise performance measure is that it disregards the settings of the uncontrollable factors.

4.2. Spline method

In order to use the information in settings of the environmental factors, we consider models with the environmental factors included. The experiments are



(a) (b)

Fig. 3. Sample mean and standard deviation of the CVaR waiting times summarized by setting for the crossed design (a) and the top-down design (b). The natural grouping is indicated by symbols; x represents the group with low sample mean and sample standard deviation, o the middle group and $+$ the group with high mean and standard deviation. (1) corresponds to having all factors at their low levels and e.g., acd to having factors A, C and D at their high level as described in Montgomery (2009)

analyzed following the standard techniques for factorial experiments (Montgomery, 2009) for the controllable factors A-D, while the uncontrollable/environmental factors are handled differently. Since little knowledge is available in advance, and to make the analysis as robust as possible, we use a generalized additive model (Hastie and Tibshirani, 1990; Wood, 2003, 2006) to estimate the impact of the environmental factors on the CVaR waiting times. By using the generalized additive model framework, the environmental effects are estimated in a non-parametric fashion. The effect of each uncontrollable factor on the output is fitted by penalized regression splines ($f_j(x_{e_j})$ in equation (4)), which implies that the model covers the range from simple regression lines to complex non-linear functions. Combining the two model parts gives the overall model as

$$\begin{aligned} E(CVaR) &= \beta_0 + \sum_{j=1}^{4} \beta_j x_j + \sum_{j=1}^{3} \sum_{k=j+1}^{4} \beta_{j,k} x_j x_k + \sum_{j=1}^{2} \sum_{k=j+1}^{3} \sum_{l=k+1}^{4} \beta_{j,k,l} x_j x_k x_l \\ &+ \beta_{1,2,3,4} x_1 x_2 x_3 x_4 + \sum_{j=1}^{8} f_j(x_{e_j}) \end{aligned} \tag{4}$$

where x_{e_j} is the j'th environmental factor, f_j its smooth function and x_1 corresponds to factor A, $x_1 x_2$ to the interaction between factors A and B, etc. The four controllable factors are all coded as -1 and 1 for the low and high levels, respectively.

In terms of the controllable factors the significant effects are the main effects of factors A, C and D in both designs. Reducing the model to having only the significant controllable factors together with the uncontrollable factors leads to insignificant increases in the residual deviance with $p = 0.30$ and $p = 0.23$ for the crossed and the top-down design, respectively. Table 4.2 summarizes the

Table 4. Significant parametric effects for crossed and top-down designs, where β_0 corresponds to the intercept, β_1 is the effect of anesthesiologists, β_3 the effect of recovery beds and β_4 the effect of operating days

	Estimate (S.E)	
Parameter	Crossed	Top-down
$\hat{\beta}_0$	70.37(0.27)	70.29(0.26)
$\hat{\beta}_1$	−3.60(0.27)	−3.69(0.25)
$\hat{\beta}_3$	−2.33(0.27)	−1.95(0.28)
$\hat{\beta}_4$	−0.60(0.27)	−0.90(0.27)

parametric effects and it can be seen that the estimates coincide. Furthermore the optimal strategy is to increase the number of anesthesiologists and recovery beds while having a week with four operating days. The number of porters is seen to have an insignificant impact on the CVaR waiting time.

The difference between the top-down design and the crossed design is, however, substantial in terms of estimating the significant environmental factors. The crossed design suggests that only the environmental factor related to occupancy of the recovery beds is significant, and this is only borderline ($p = 0.07$ as the highest p-value). In contrast, the top-down design identifies three highly significant factors; the acute arrival rate, the occupancy of the recovery beds and the occupancy of the anesthesiologist ($p \leq 0.02$). The effects of the significant environmental factors in the top-down design are shown in Figure 4. The corresponding plots for the crossed design are shown in the lower part of Figure 4 as reference, which shows that only the environmental factor related to occupancy of the recovery beds is borderline significant.

The crossed design is based on only ten environmental settings, which implies that the corresponding estimated effects become highly uncertain. In contrast the effects estimated with the top-down design are estimated with much higher certainty. From Figure 4, for example, it can also be seen that as the acute arrivals are increased, the waiting time increases. Likewise the effects of having less access to recovery beds and anesthesiologists (higher occupancy) cause significant increases in the waiting time. The impact on the waiting time is seen to be most pronounced for occupancy of the recovery beds and the anesthesiologists.

By combining the parametric and smoothed functions it is seen that factors A (the anesthesiologists) and C (the recovery beds) are the the most important factors; they have the largest estimated effects and moreover the environmental effects related to factors A and C (the occupancy of the anesthesiologist and the occupancy of the recovery beds) are also highly significant.

In order to further investigate the significant uncontrollable factors in the top-down design, we include interaction terms between the controllable factors recovery beds and anesthesiologists and their associated uncontrollable factors, occupancy of recovery beds and occupancy of anesthesiologists, in the reduced model. We restrict ourselves to considering only these interactions because there is a direct connection between the controllable and uncontrollable factors for these two factors. The inclusion of interactions between controllable and uncontrollable factors is also recommended by for example Bursztyn

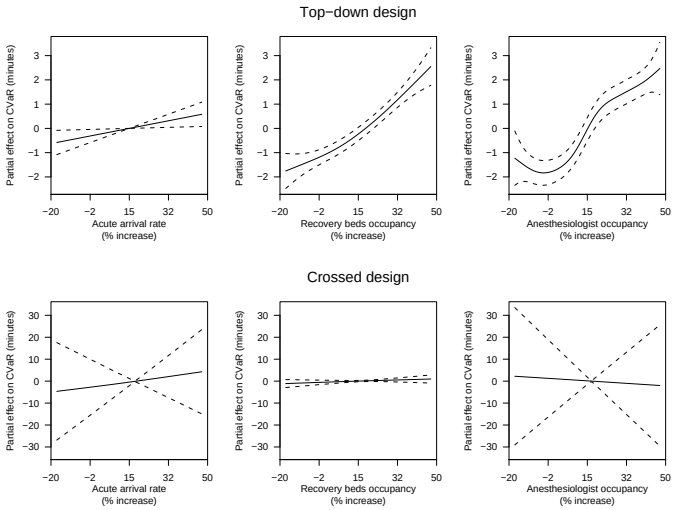


Fig. 4. Significant environmental factors. The solid lines indicate the mean effect and the dashed lines the uncertainty of the mean effect. The x axis is measured in % corresponding to the 20 % better and 50 % worse scenario range used in the experiment plan for the environmental factors

and Steinberg (2006) and Myers et al. (2009) for physical experimentation. This gives the following model

$$\begin{aligned} E(CVaR) &= \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + f_1(x_{e_2}) + f_2(x_{e_1}) z_{(-1)1} \\ &+ f_3(x_{e_1}) z_{11} + f_4(x_{e_3}) z_{(-1)3} + f_5(x_{e_3}) z_{13} \end{aligned} \tag{5}$$

where x_1, x_3 and x_4 are as defined in equation (4), x_{e_2} is the acute inter- arrival rate, x_{e_1} is the occupancy of the anesthesiologists and x_{e_3} is the occupancy of the recovery beds and z_{ij} is an indicator variable for whether controllable factor x_j has level i.

Estimating the model in equation (5) shows that two environmental factors interact with their respective controllable factors (the occupancy of the recovery beds and the anesthesiologists) in the top-down design. Figure 5 shows the interactions, which can be seen to be a steeper linear effect at the low level of the recovery beds (factor C) compared to the high level for the occupancy of the recovery beds. For the anesthesiologist resource, it can be seen that, at the low level, the estimated effect is linear and, at the high level, an S-shaped curve is seen, the latter indicating that the setting is robust up to a certain level, as we initially observe a flat curve. For the crossed design the occupancy of the anesthesiologist is insignificant and the occupancy of the recovery beds is only (borderline) significant at the low level for the number of recovery beds.

From Figure 5 it can be seen that the analysis of the top-down experiment suggests that the system is much more robust in terms of the CVaR with high levels of recovery beds and anesthesiologists. However this is not picked up by the crossed design, for which the analysis shows a borderline significant interaction between occupancy of the recovery beds and the number of recov-

ery beds. Moreover, the interaction for anesthesiologists and occupancy of the anesthesiologists is seen to be insignificant.

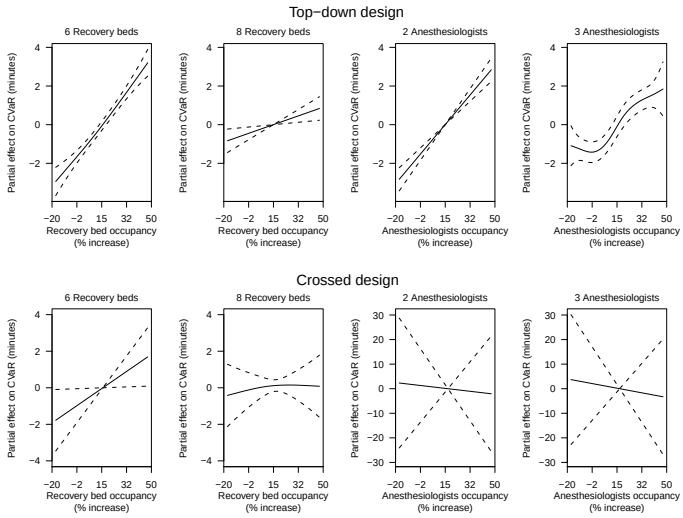


Fig. 5. Estimated interactions between environmental and controllable factors. The solid lines are the estimated mean effects and the dashed lines indicate the uncertainty of the mean effects

5. Discussion

This study is based on an application of a discrete event simulation model of a hospital unit. In healthcare applications in general, it is desirable that the final solutions are robust to changes in the uncontrollable factors. In the proposed design a large set of combinations of the uncontrollable factor settings is achieved by using only a limited number of runs for each controllable factor setting. This is done by using a different set of uncontrollable factor settings for each controllable factor setting. Moreover the subplots are selected so that, when considered together, they provide uniform coverage of the design space. One restriction in the design method is the number of subplots which needs to be the same in all whole plots. Unbalanced designs may also be of interest, but this would require a more general construction method.

Qian et al. (2009) and Qian et al. (2009) propose designs where a high-accuracy experiment is nested within a low-accuracy experiment. The main idea is to construct two experiments, where the smaller one is nested in the complete design. Qian et al. use this for cases where two computer codes for the same problem are available; one slow but accurate and one fast but less accurate. Thus the experimenter wants to run fewer experiments with the slow code but more using the fast code. Qian and Wu (2008) integrate the information in the two experiments using a Bayesian hierarchical model. The model is primarily built on the low-accuracy experiment, whereas the high-accuracy experiments are used to calibrate and correct the model such that it fits the high-accuracy code. Calibration is done on points that the

two experiments have in common. In a recent paper Qian and Wu (2009) consider a slice space-filling design, which is based on latin hypercubes from a customized orthogonal array for the quantitative factors. The overall design is then sliced into subdesigns corresponding to the setting of the qualitative factor settings.

Rennen et al. (2009) consider nested maximin latin hypercube designs. They consider the nested design useful in the dual experiments described by Qian et al., but also for developing training and test data sets and for sequential experimentation. For the development of the training and test data sets, the design procedure can provide the experimenter with a space-filling (with respect to the max-min criteria) design for the training data and a larger test data set, which, together with the training data set, is also space-filling. Similarly for sequential experimentation, a small space-filling experiment is initially run and then potentially expanded with further experimentation by evaluating the complete design, which once again, together with the initial design, also forms a space-filling design. Sequential sampling is also considered by for example van Beers and Kleijnen (2008, 2003) and Kleijnen and van Beers (2004) for metamodeling with kriging. Sequential sampling with controllable and uncontrollable factors is an interesting strategy for future research but beyond the scope of the current work.

In the case study presented in section 4 it is shown that the top-down design is better suited for estimating the environmental effects compared to the crossed design. The estimated parametric effects in the two designs coincide in terms of the three factors of major importance. It was shown that the crossed design overlooked some of the important environmental effects, since the coverage of the environmental factor space was worse. More importantly, the crossed design overlooked significant interactions between controllable and uncontrollable factors. Identifying these interactions is crucial to being able to set the system in a robust operating mode. Thus, the significantly better coverage of the environmental factor space implies that analysis based on the top-down approach is less likely to overlook important effects of the uncontrollable factors as well as important interactions between controllable and uncontrollable factors.

In this paper we consider spline models for analyzing the output from the simulation model. In the deterministic computer experiments literature the kriging (DACE) model is often used (Santner et al., 2003; Sacks et al., 1989). For simulation models Kleijnen (2008, 2009) and Ankenman et al. (2010) consider kriging for stochastic simulation models. Kleijnen (2008, 2009) uses bootstrap methods for estimating the uncertainty around the kriging predictor, whereas Ankenman et al. (2010) expand the usual kriging model with an extra stochastic component corresponding to the variation for replications. These methods may be relevant for the type of application presented in this paper. One limitation of the above methods is that the factors are considered to be continuous, which is not the case for the controllable factors in our study.

6. Conclusion

In this study, a methodology for the design of uniformly distributed experiments for simulation experimentation in the presence of both controllable and uncontrollable factors is introduced. The methodology ensures that the uncontrollable factor settings in the combined design for the uncontrollable factors are uniform, while keeping an acceptable level of uniformity of the subplots for each controllable factor setting.

The proposed methodology is primarily based on Euclidean distances. Therefore the method can be used in designs with many uncontrollable/environmental factors. Our results show that the method is applicable to designs with two to 19 uncontrollable factors. Because the methodology is based on distances, increasing the number of factors may be possible, although the sparsity of experiments in the design space may become an issue.

For our case study it was shown that the effects of the uncontrollable factors, together with the interaction between controllable and uncontrollable factors, were significantly better estimated with the proposed design compared to a crossed design. The crossed experiment overlooked the important interactions between controllable and uncontrollable factors, and these are important for making the system robust. This also implies that the uncontrollable effects are better understood with the top-down design. Moreover, since the uncontrollable factor space is better covered with the top-down approach, the reliability of the results is higher compared to a crossed design. The results in terms of the controllable part of the model were seen to be the same in both designs, which implies that the benefit of the proposed design is primarily related to the extended coverage of the uncontrollable factor space.

In future work we focus on the analysis part; i.e., applying the Kriging model on the output from the proposed design. The Kriging model is very popular in simulation and an useful extension to the Kriging model will be to incorporate the uncontrollable/controllable factor framework discussed in this paper.

References

Alexander, S., T. Coleman, and Y. Li (2006). Minimizing cvar and var for a portfolio of derivatives. *Journal of Banking and Finance 30*(2), 583–605.

Ankenman, B. E., B. L. Nelson, and J. Staum (2010). Stochastic kriging for simulation metamodeling. *Operations Research*. forthcoming.

Bielen, F. and N. Demoulin (2007). Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality 17*(2), 174–193.

Bursztyn, D. and D. Steinberg (2006, july). *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, Chapter Screening Experiments for Dispersion Effects, pp. 21–47. Springer New York. Editors: A. Dean and S. Lewis.

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2008). Designing simulation experiments with controllable and uncontrollable factors. In *Proceedings of the 2008 Winter Simulation Conference, Miami, FL, 2008.*

Dehlendorff, C., M. Kulahci, S. Merser, and K. K. Andersen (2010). Conditional value of risk as a waiting time measure in simulations of an orthopedic surgery. *Quality Technology and Quantitative Management*. To appear.

Fang, K.-T., R. Li, and A. Sudjianto (2006). *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC.

Fang, K.-T., X. Lu, and P. Winker (2003). Lower bounds for centered and wrap-around l2-discrepancies and construction of uniform designs by threshold accepting. *Journal of Complexity 19*(5), 692–711.

Fang, K.-T. and C.-X. Ma (2001). Wrap-around l2-discrepancy of random sampling, latin hypercube and uniform designs. *Journal of Complexity 17*(4), 608–624.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hickernell, F. (1998a). *Random and Quasi-Random Point Sets*, Chapter Lattice rules: How well do they measure up?, pp. 106–166. Springer-Verlag, New York.

Hickernell, F. J. (1998b). A generalized discrepancy and quadrature error bound. *Mathematics of Computation 67*(221), 299–322.

Kibzun, A. and E. Kuznetsov (2003). Comparison of var and cvar criteria. *Automation and Remote Control 64*(7), 153–164.

Kleijnen, J. and W. van Beers (2004). Application-driven sequential designs for simulation experiments: Kriging meta-modeling. *Journal of the Operational Research Society 55*, 876–883.

Kleijnen, J. P. (2008). *Design and Analysis of Simulation Experiments*. Springer.

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research 192*(3), 707–716.

Krahl, D. (2002). The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*, pp. 205–213.

Montgomery, D. C. (2009). *Design and Analysis of Experiments* (7th ed.). John Wiley and Sons, Inc.

Myers, R., D. Montgomery, and C. Anderson-Cook (2009). *Response surface methodology: process and product optimization using designed experiments* (3rd ed.). Wiley, New York.

Qian, P. Z. G., M. Ai, and C. F. J. Wu (2009). Construction of nested space-filling designs. *The Annals of Statistics 37*(6A), 3616–3643. DOI: 10.1214/09-AOS690.

Qian, P. Z. G., B. Tang, and C. J. Wu (2009). Nested space-filling designs for computer experiments with two levels of accuracy. *Statistica Sinica 19*, 287–300.

Qian, P. Z. G. and C. F. J. Wu (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics 50*(2), 192–204.

Qian, P. Z. G. and C. F. J. Wu (2009). Sliced space-filling designs. *Biometrika 96*(4), 945–956.

Rennen, G., B. Husslage, E. R. van Dam, and D. den Hertog (2009). Nested maximin latin hypercube designs. *CentER Discussion Paper* (2009-06).

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical Science 4*(4), 409–423.

Sanchez, S. M. (2000). Robust design: Seeking the best of all possible worlds. In *Proceedings of the 2000 Winter Simulation Conference*, pp. 69–76.

Santner, T. J., B. J. Williams, and W. I. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer.

Taguchi, G. (1987). *System of experimental design, volumes 1 and 2*. UNIPUB/Krauss International, White Plains, New York.

van Beers, W. and J. Kleijnen (2003). Kriging for interpolation in random simulation. *Journal of the Operational Research Society 54*, 255–262.

van Beers, W. C. and J. P. Kleijnen (2008). Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research 186*(3), 1099–1113.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B 65*, 95–114.

Wood, S. (2006). *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC.

Paper D

Analysis of Computer Experiments with Multiple Noise Sources (European Network for Business and Industrial Statistics)

Conference paper published in Proceedings of ENBIS8, Athens 2008 (non peer-reviewed)

Analysis of Computer Experiments with Multiple Noise Sources

Christian Dehlendorff Murat Kulahci Klaus Kaae Andersen

1 Introduction

In the classic computer experiments analysis the output from the computer model is deterministic [18, 16]. For deterministic output a natural requirement is that the predictor interpolates the data, since the output is observed without noise. Kriging [8, 10] is an often used modeling technique, where interpolation is incorporated by the specification of a covariance function depending on distances to the observed data.

The focus in computer experiments is often the deterministic/fixed effects, i.e. which parameter settings yield the best outcomes. However, some applications includes factors that are uncontrollable in the sense that they can not be controlled in the physical system. Such uncontrollable factors could for example be the customer arrival frequency in a grocery store or the room temperature in a laboratory. The levels of the uncontrollable factors can not be decided by experimenter and the factors therefore need to be treated differently in the analysis. The analysis of uncontrollable factors is the focus of this paper.

Kleijnen [7] considers simulation models as a special class of computer models, which typically includes one (or more) stochastic part(s). The sources of variation are the seed controlling the random number generator and the uncontrollable factors included in the model to account for environmental variations. The variation from varying the seed in a simulation model arises from the embedded stochastic components such as queues and activities and can be considered as corresponding to experimental error in a physical experiment.

The second type of variation in simulation models is coming from changes in the uncontrollable factors. The uncontrollable factors are settings that, although in the simulation model are fixed, can not be controlled in the physical system. To mimic the uncertainty from the environment the settings of the uncontrollable factors are varied (see section 4). The random effects associated with the uncontrollable factors are important for the robustness [17]. Often the functional relationship between the uncontrollable factors and the outcome is left

unspecified and considered merely as a source of variation. Kleijnen [7] suggests for robustness analysis to summarize the mean and the variability for each controllable factor settings and model them by two separate second order polynomials.

Another approach would be to model the functional relationship between the outcome and the uncontrollable factor. This may unveil which uncontrollable factors are important. Moreover, if the uncertainty of an important uncontrollable factor can be improved by e.g. quality improvements the functional relationship could quantify the gain by doing so.

In this paper the sources of variation are quantified by means of a linear mixed effects model to separate the variation into a component corresponding to changing the uncontrollable factor settings and a component corresponding to the seed. Additionally, a generalized additive model is introduced as an easy to use tool for modeling the functional relationship between the outcome and the uncontrollable factors, i.e. model the variance components from the linear mixed effects model.

2 The case-study

The system considered in this paper is a discrete event simulation model of an orthopaedic surgical unit. The discrete event simulation model describes the individual patient's flow through the unit (illustrated in Figure 1) and is developed in collaboration with medical staff at Gentofte University Hospital in Copenhagen. The unit undertakes both acute and elective (planned) surgery and performs more than 4,600 operative procedures a year. While patients come from various wards throughout the hospital, the main sources of incoming patients are the four orthopedic wards and the emergency care unit.

The simulation model includes two sources of noise coming from changes in the uncontrollable factors (a.k.a. environmental factors in physical experimentation) and from changes in the seed controlling the random number generation process embedded in the simulation model. The uncontrollable factors are for example the arrival rate of acute patients and cleaning time of the operating rooms. Moreover, a set of controllable factors, for example the number of operating rooms and the number of surgeons, is included. Typical outcomes are waiting times, patient throughput (the total number of patients treated) and the amount of overtime used on elective operations. The simulation model is implemented in Extend [9] and controlled from a Microsoft Excel spreadsheet with a Visual Basic for applications script.

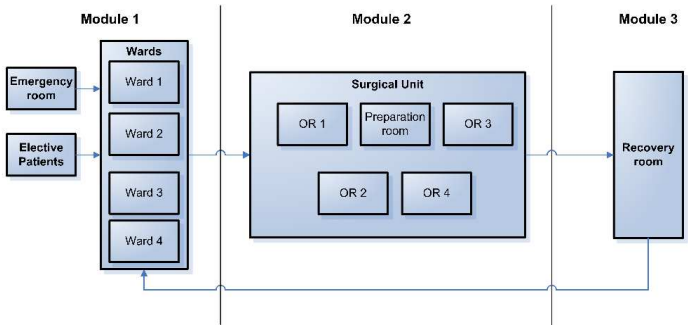


Figure 1: Basic layout of surgical unit. The patient flow is from left to right.

2.1 Performance measures

The performance measures considered for the simulation model are the total throughput (TT), the percentage of elective patients treated outside regular hours (EOUT) and the extent of long waiting times. Often the long waiting times are the most important ones since they from the patient perspective tend to be the most bothersome [1]. The waiting time distribution for the case-study is highly right-skewed with a minimum of 0 minutes, a mean of 28 minutes, a 95 % quantile of 51 minutes and a maximum of 140 minutes, which shows that long waiting times are present.

We suggest measuring the extent of long waiting times by the Conditional Value at Risk (CVaR) measure. The measure originates from economics as an extension of Value at Risk (VaR) [15, 5, 6]. Both measures quantify a distribution of losses in e.g. portfolio management with a single statistic. For the set of waiting times $T_x = \{t_{x1}, \ldots, t_{xN}\}$ from the x'th run, $CVaR_\alpha(T_x)$ is defined as the expected value of the α-tail distribution of T_x [15], i.e.

$$CVaR_\alpha(T_x) = \frac{\left(\frac{i_\alpha}{N} - \alpha\right) t_{xi_\alpha} + \sum_{i=i_\alpha+1}^{N} \frac{t_{xi}}{N}}{1-\alpha} \tag{1}$$

with $t_{x1} \leq t_{x2} \leq \cdots \leq t_{xN}$, i_α is the index satisfying $\frac{i_\alpha}{N} \geq \alpha > \frac{i_\alpha - 1}{N}$. t_{i_α} is the α-quantile and in economics denoted the Value at Risk (VaR). CVaR can be seen as a compromise between the average waiting time ($\alpha = 0$) and the maximum waiting time ($1 - 1/N < \alpha < 1$), where α reflects the weight of the longest waiting times in the measure. In the following $\alpha = 0.95$ is used corresponding to that CVaR is the average of the 5% longest waiting times.

TT and EOUT are quality measures that are required to fulfill the quality constraints

1. At least the same number of patients treated compared to the reference setting

2. At least the same percentage of elective patients treated outside regular hours compared to the reference setting

where the reference setting corresponds to the current setting (see section 4), i.e. corresponding to the performance under the current resource allocation at the department. The requirements are constraints that ensures that a low CVaR is not obtained by treating fewer patients or by treating more patients outside regular hours.

Two main questions that involves treatment of the uncontrollable factors are addressed in this paper

1. How big are the variations in long waiting times?
   (a) from which sources do they arise?
   (b) which uncontrollable factors are influential?
2. Can the risk of not meeting the quality requirements for the total patient throughput and the extent of overtime be minimized?

3 Model

Models that interpolate the data are not a requirement for non-deterministic output, which imply that the kriging framework looses its intuitive appeal. Our case study furthermore complicates the analysis, since most controllable factors are discrete. Moreover, the presence of uncontrollable factors implies that the factors fall in two groups with different interpretations. We focus on the uncontrollable factors and treat the controllable factor settings as a single factor. As a starting point a linear model is considered

$$y(x_{c_i}, x_{e_j}, s_k) = \beta_i + \epsilon_k \tag{2}$$

where β_i is the effect of controllable setting i and $\epsilon_k \sim N(0, \sigma_\epsilon^2)$ the residual variation. This model has parameters for each contpollable factor setting and a single erros term for the variation corresponding to the seed and the uncontrollable factor settings.

The linear model estimates the variations related to the uncontrollable factors and the seed separately. To target both types of variations explicitly a linear mixed effects model (LME) [13] is proposed. The LME is formulated such that it quantifies the two sources of variation, i.e.

$$y(x_{b_i}, x_{e_j}, s_k) = \beta_i + E_j + S_k \tag{3}$$

where β_i is the effect of controllable setting i, $E_j \sim N(0, \sigma_E^2)$ is the variation from the varying uncontrollable factor settings and $S_k \sim N(0, \sigma_S^2)$ the variation corresponding to

the seed. The variation corresponding to changes in the uncontrollable factors is modeled by considering the j'th uncontrollable factor setting's effect as random $E_j \sim N(0, \sigma_E^2)$. The remaining variation is contained in the S_k's. In gage R&R terminology the seed variation, σ_S^2, corresponds to the repeatability and the total variance (the σ_ϵ^2 in the linear model), $\sigma_T^2 = \sigma_E^2 + \sigma_S^2$, to the reproducibility [12].

An alternative approach is to model the functional relationships between y and the uncontrollable factors. This functional relationship can straight forward be estimated with a Generalized Additive Model (GAM) [19]. The GAM models the functional relationship by a sum of additive smooth functions

$$y(x_{c_i}, x_{e_j}, s_k) = \beta_i + \sum_{l=1}^{m} f_l(\tilde{x}_{e_j}^l) + S_k \tag{4}$$

with $\tilde{x}_{e_j}^l$ being the j'th setting for the l'th uncontrollable factor and $S_k \sim N(0, \sigma_S^2)$ the residual or seed term. f_l is a spline based smooth function with the smoothness determined by a penalty term. By estimating the functional relationship between the uncontrollable factors and the outcome, the factors most important to control (if possible) are identified. This could be the basis for focused strategies for reducing the environmental variations, i.e. corresponding to reducing σ_E^2 in the LME.

The risk of not fulfilling the quality requirements can also be analyzed within the GAM framework. For the output y_q and the quality requirement c_q, the outcome is binary, $I(y_q < c_q)$. A GAM with a binomial distribution family is considered and the linear predictor is given as

$$E\left[\log\left(\frac{p}{1-p}\right)_{ij}\right] = \beta_i + \sum_{l=1}^{m} f_l(\tilde{x}_{e_j}^l) \tag{5}$$

where p is the risk of not meeting the requirements.

The advantage of using the GAM framework is that the interpretation of the smoothed functions is intuitive and can for example be presented graphically to the medical staff. Moreover, the GAM does not impose a parametric form on the functional form (besides the additivity), which imply that the data drives the analysis. Another advantage is that the controllable factor settings are corrected by the levels of the uncontrollable factors. The disadvantage of the GAM framework is the additivity assumption, which in this paper implies that only marginal effects are considered. It is possible to expand the GAM to include functions of more than one variable and interactions with e.g. controllable factors, which potentially could involve rather complex meta models. Moreover, GAM modeling methods are freely available in statistical software [19, 14].

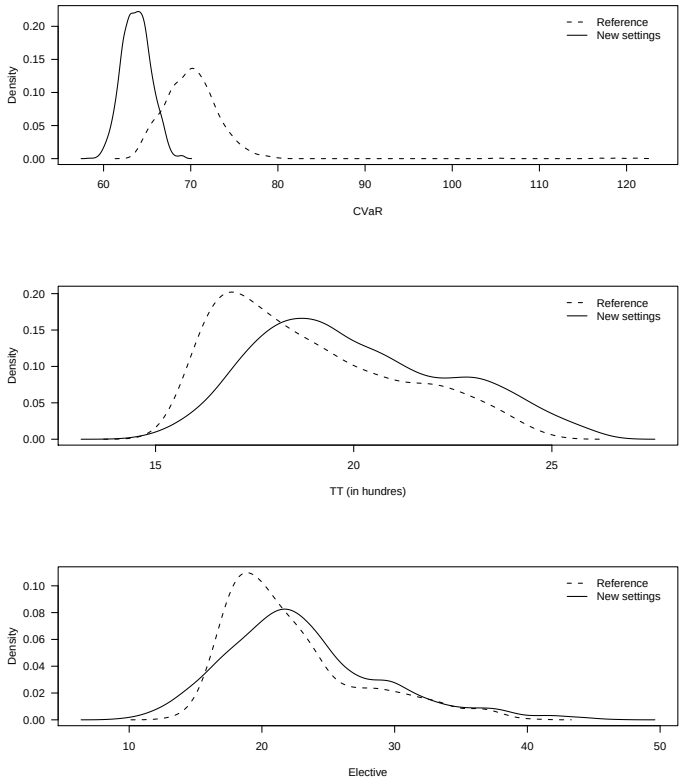


Figure 2: Estimated densities for CVaR (top), TT (middle) and EOUT (bottom) for reference design (dashed lines) and new settings (solid lines)

4 Data

In the remaining part of the paper output from the simulation model is considered. The average run time for simulating 6 months operation (with one week of warm-up) is around 7 minutes. For each run the system's performance is summarized in a set of measures, e.g. the total patient throughput, the percentage of elective patients treated outside regular hours and the CVaR waiting time. Two sets of data are considered:

1. 1 controllable factor setting corresponding to the current setting with
   (a) 400 different uncontrollable factor settings chosen such that the ranges of the 8 uncontrollable factors are uniformly covered with respect to the wrap around L_2 discrepancy [4, 3]
   (b) 2 repetitions with different seeds for each uncontrollable factor setting, i.e. a total of $N = 800$ runs
2. 20 different controllable factor settings, which were found in a pilot study with the objective of finding good settings in terms of reducing the predicted CVaR waiting time while maintaining the performance on TT and EOUT
   (a) each controllable setting was assigned 20 different uncontrollable factor settings by splitting a 400 run uniform design with 8 factors into 20 sub designs
   (b) sub designs were generated such that the wrap around L_2 discrepancy uniformity criteria was minimized
   (c) 5 repetitions with different seeds for each uncontrollable and controllable factor combination, i.e. a total of $N = 2000$ runs

The analysis here is concerned about the second experiment if not stated otherwise, whereas the first experiment serves as reference. The outputs from the two simulation experiments are shown in Figure 2. The CVaR waiting times are the averages of the 5 % longest waiting times in each run corresponding to the 90-100 longest waiting times. The potential range is from the 95 % quantile (51 minutes) to the maximal waiting time (140 minutes). However, as the waiting time distribution is right skewed the CVaR-values tend to be in the range from 55 to 80 minutes with the exception of 4 observations in the reference experiment.

5 Results

Figure 2 shows the CVaR waiting times for the 20 new settings and the current settings. It is seen that the waiting times for the new settings are lower compared to the current setup. Furthermore, the coefficient of variation is lower for CVaR for the new settings (CV=2.58 %)

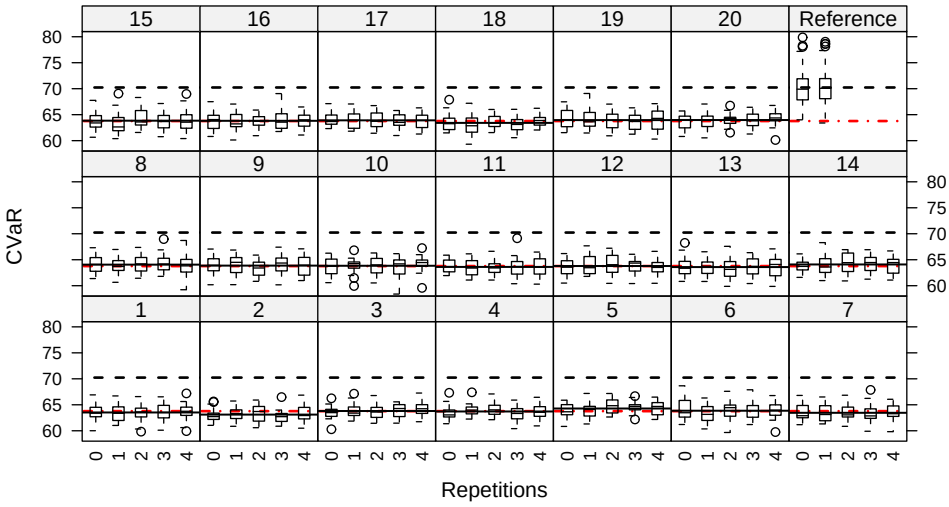


Figure 3: Box plots of CVaR for the 20 new setting (the panels labeled 1 to 20 above the panel) and the reference setting (labeled Reference). The dashed line corresponds to the overall mean in the reference design, the dot-dashed to the overall mean of the 20 new settings and the solid lines to the individual setting means. Note that the box plot for the reference has been cut off at 80, which imply that 4 observations are missing see section 5

Table 1: Variance components, overall means and adjusted R^2 for the 20 suggested settings and the reference scenario, respectively. † Linear regression model. ⋆ The outcome has been square root transformed. * The linear model corresponds to the null-model since only one controllable factor setting is present.

		New settings			
Outcome	Model	σ^2_E	σ^2_S	R^2_a	μ
CVaR	LM†	-	1.63^2	0.02	
	LME	1.17^2	1.16^2	-	63.77
	GAM	-	1.15^2	0.51	
EOUT⋆	LM	-	0.50^2	0.33	
	LME	0.48^2	0.18^2	-	4.77
	GAM	-	0.18^2	0.91	
TT	LM	-	223.50^2	0.18	
	LME	224.01^2	42.41^2	-	2005.45
	GAM	-	42.74^2	0.97	
		Reference scenario			
		σ^2_E	σ^2_S	R^2_a	μ
CVaR	LM†	-	4.34^2	0*	
	LME	2.19^2	3.74^2	-	70.23
	GAM	-	3.69^2	0.28	
EOUT⋆	LM	-	0.52^2	0*	
	LME	0.48^2	0.19^2	-	4.68
	GAM	-	0.18^2	0.88	
TT	LM	-	226.80^2	0*	
	LME	222.60^2	43.73^2		1888.92
	GAM	-	41.70^2	0.97	

compared to the reference (CV=6.18 %), TT (CV_{old} = 12.01 % and CV_{new} = 12.23 %) and EOUT [1] (CV_{old} = 11.03 % and CV_{new} = 12.81 %). The increase in the CV in the reference scenario for the CVaR waiting times is caused by the right skewed distribution with observations ranging from 63.40 to 121.17 minutes. Without the 4 largest observations the CV reduces to 4.13 %, i.e. still considerable higher. The overall mean of the CVaR was estimated to 63.77 and 70.23 minutes for the new settings and the reference setting, respectively.

The CVaR waiting times from the two experiments are summarized by box plots in Figure 3. From the figure it is seen that most of the variation in the new settings can be attributed to variations in the uncontrollable factors and the seed. The controllable factor setting means are seen to be distributed closely. The linear model considered in Table 1 does indicate significant differences between the 20 new settings with setting 2 being the setting with the lowest CVaR waiting time. Furthermore, the variances of for the residuals by controllable setting show evidence of being heterogeneous ($p = 0.005$ for Bartlett's test of variance homogeneity). Moreover, Figure 3 indicates that the reference setting is more sensitive to the uncontrollable factor settings compared to the new settings.

5.1 LME

The REML variance components in the LME of the CVaR for the new settings are summarized in Table 1. The two components for CVaR are seen to be comparable in size and a bootstrapped 95 % confidence band [2] for the intraclass correlation [11] gives $0.46 \leq \frac{\sigma_E^2}{\sigma_E^2+\sigma_S^2} \leq 0.55$. For the reference setup the variance components of the CVaR are seen to be significantly larger, which shows that not only is the current setup inferior to the proposed setups it also tends to be more sensitive to changes in the uncontrollable factors and the seed.

Figure 2 shows that the CVaR waiting time in the reference scenario is a highly right skewed distribution with 4 runs with values above 100, whereas the remainder of the runs are contained in the interval [63.40; 79.87]. The 4 observations furthermore violate the model assumptions: $B_j \sim N(0, \sigma_E^2)$ and $S_k \sim N(0, \sigma_S^2)$. Omitting the observations from the analysis gives $\sigma_E^2 = 2.36^2$ and $\sigma_S^2 = 1.67^2$, which is seen to increase σ_E^2 and decrease σ_S^2 (the average decreases from 70.23 to 70.00). The intraclass correlations before and after removing the 4 observations are 0.26 and 0.67 corresponding to the difference between seeds is significantly smaller after the removal. The diagnostics after omitting the observations do not indicate problems with the model assumptions. The size of the variance components for TT and EOUT are seen to be equivalent for the two experiments. The analysis shows that the old setting is most sensitive to changes in the uncontrollable factors.

5.2 GAM

To identify the important uncontrollable factors a GAM model with smooth functions for each of the 8 uncontrollable factor and a parameter corresponding to each of the 20 controllable factor settings is fitted. The GAM shows that 4 uncontrollable factors are significant associated with the CVaR waiting times while the remaining 4 uncontrollable factors seem not to be related to the CVaR waiting time. The significant factors are the incoming rate of acute patients and the amount of time the anesthesiologists, porters and the recovery beds are occupied by other activities.

The estimated functional forms of the 4 significant factors are illustrated in Figure 4. The curves fitted for each of the 5 repetitions for the new settings show that the functional form is consistent from one repetition to the next. It is from Table 1 seen that the residual variation is estimated to $\sigma_S^2 = 1.15^2$, which is seen to match the component from the LME. This indicates that no information is lost by requiring the smooth functions to be additive. Moreover, the adjusted R^2s show that the benefits of including the uncontrollable factors are significant with absolute improvements in R^2 by 0.50 or more compared to the linear

[1]Square root transformed for symmetry and for consistency with Table 1

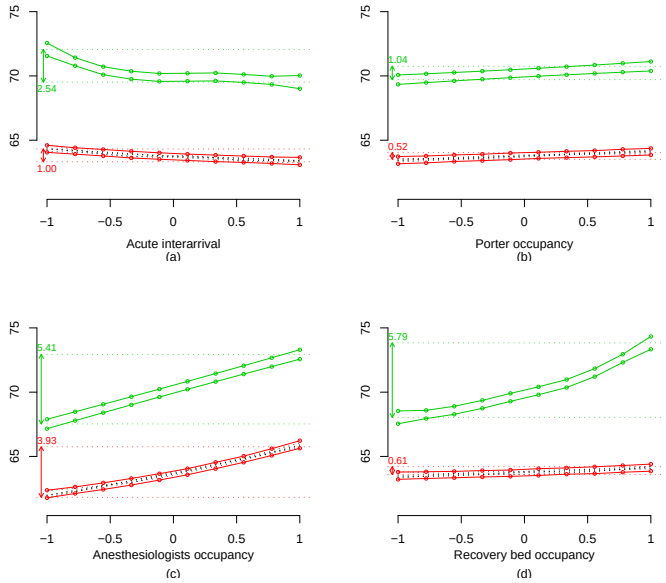


Figure 4: Significant uncontrollable factors. The two top curves in each of the 4 sub figures correspond to the 95 % confidence limits in the reference design. The bottom curves consists of two solid curves corresponding to a 95 % confidence limits in a model with all 5 repetitions included and 5 dashed curves corresponding to each of the repetitions.

model.

Figure 4 shows that the same functional relationships are present for the uncontrollable factors in the reference design except for the occupancy of the recovery beds. The occupancy of the recovery beds has a steeper increase in CVaR in the reference setting compared to the new settings, which is likely to be caused by the fact that fewer beds are available in the reference setting. The smoothed curves for the occupancy of the recovery beds show that the new settings are more robust against variations in this factor.

5.3 Risk profiles

The risk profiles of CVaR, TT and EOUT as function of the controllable settings are shown in Figure 5. The risks are defined as the risk of not fulfilling the quality requirements defined in section 2.1 after adjusting for the uncontrollable factor settings. In addition to the already defined requirements, it is for CVaR waiting time required that the new settings

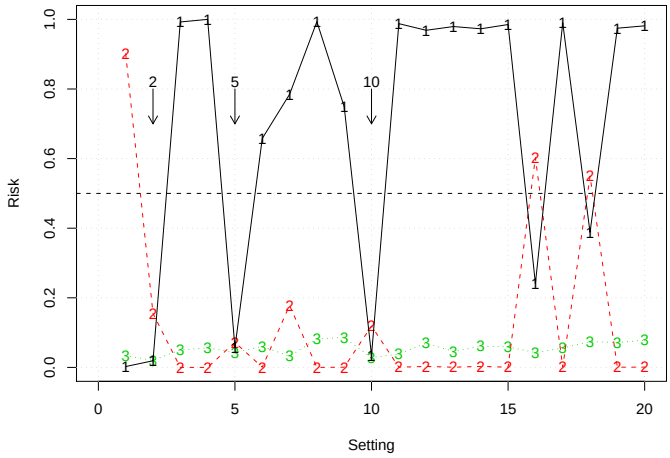


Figure 5: Risk profiles for CVaR (dotted line marked *3*), TT (dashed line marked *2*) and EOUT (solid line marked *1*). Arrows indicate settings with risks lower than 0.5 (marked by dashed line) for TT and EOUT.

have a lower CVaR-value than the 5 % quantile in the reference setting (65.43 minutes). Table 1 shows that the performance in both mean value and variance components is similar for TT and EOUT with the new setting compared to the current setting. On average the TT is better (higher) with the new settings, whereas EOUT is worse (higher).

The risks are estimated with a GAM, which models the 8 uncontrollable factors with smooth functions and the controllable factors settings as one factor. For the risks corresponding to TT and EOUT, it is seen that settings 2, 5 and 10 perform well for both measures. It can also be seen that the TT and EOUT risks are negatively correlated (Spearman's rho: -0.89), i.e. that lowering the risk of treating to few patients increases the risk of treating more elective patients outside regular hours.

The risk of exceeding the 5 % quantile in the CVaR distribution for the reference scenario is lowest for setting 2, which coincide with Figure 3. The 3 solutions are quite similar, i.e. they operate with 4 operating days, 4 operating rooms and an increase in elective patients by 2 per day. The 3 proposed settings use more resources compared to the current setup with the lowest additional costs for setting 5. It is seen that all 3 suggested settings on average fulfill the requirements in more than 80 % of the runs.

6 Conclusion

The main contribution in this paper was the analysis of the simulation model, which involved two sources of variation. The results showed that the variations in the CVaR waiting time with a linear mixed effects model could be split into two equally large variance components for the new settings, whereas the seed variance in the reference scenario was lower compared to the variance caused by changes in uncontrollable factors. The generalized additive model showed that the main source of variation for the new settings was the occupancy of the anesthesiologist. Moreover, the new settings eliminated the impact of one of the important uncontrollable factors with the reference setting.

The use of the linear mixed effects model gave insight to the extent of uncontrollable variation and the generalized additive model identified the most important uncontrollable factors. This may assist decision makers to construct focused strategies to control the uncontrollable factors better.

Moreover, the quality constraints were seen to be fulfilled in more than 80 % of the time for 3 specific settings. The total throughput and the CVaR waiting time criteria were the constraints most easy to fulfill. The draw back of the improvements in the CVaR waiting time was the cost of the additional resources needed. By combining cost and performance it may be possible to find solutions with a CVaR performance inferior to the new settings but at a significant lower cost while still improving the performance compared to the reference. Moreover more complex model structures may give a deeper understanding of the system.

References

[1] Frédéric Bielen and Nathalie Demoulin. Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality*, 17(2):174–193, 2007.

[2] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[3] Kai-Tai Fang, Runze Li, and Agus Sudjianto. *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC, 2006.

[4] Kai-Tai Fang and Chang-Xing Ma. Wrap-around l2-discrepancy of random sampling, latin hypercube and uniform designs. *Journal of Complexity*, 17(4):608–624, 2001.

[5] A.I. Kibzun and E.A. Kuznetsov. Comparison of var and cvar criteria. *Automation and Remote Control*, 64(7):153–164, 2003.

[6] Andrey I. Kibzun and Evgeniy A. Kuznetsov. Analysis of criteria var and cvar. *Journal of Banking & Finance*, 30(2):779–796, 2006.

[7] Jack P.C. Kleijnen. *Design and Analysis of Simulation Experiments*. Springer, 2008.

[8] Jack P.C. Kleijnen. Kriging metamodeling in simulation: a review. *European Journal of Operational Research*, 2008.

[9] David Krahl. The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*, pages 205–213, 2002.

[10] Jay D. Martin and Timothy W. Simpson. Use of kriging models to approximate deterministic computer models. *AIAA Journal*, 43(4):853–863, 2005.

[11] Douglas C. Montgomery. *Design and Analysis of Experiments*. John Wiley and Sons, Inc, 6th edition, 2005.

[12] Douglas C. Montgomery. *Introduction to Statistical Quality Control*. John Wiley & Sons, Inc., 2005.

[13] Jose Pineiro and Douglas Bates. *Mixed Effects Models in S and S-PLUS*. Springer, 2000.

[14] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[15] R. Tyrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26:1443–1471, 2002.

[16] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[17] Susan M. Sanchez. Robust design: Seeking the best of all possible worlds. In *Proceedings of the 2000 Winter Simulation Conference*, pages 69–76, 2000.

[18] Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[19] S.N. Wood. *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC, 2006.

Paper E

Analysis of Computer Experiments with Multiple Noise Sources

Published in Quality and Reliability Engineering International, Volume 26 Issue 2, March 2010, p. 147-155 (special issue for ENBIS8)

Analysis of Computer Experiments with Multiple Noise Sources

Christian Dehlendorff, Murat Kulahci and Klaus Kaae Andersen

Abstract

In this paper we present a modeling framework for analyzing computer models with two types of variation. The paper is based on a case study of an orthopedic surgical unit, which has both controllable and uncontrollable factors. Our results show that this structure of variation can be modeled effectively with linear mixed effects models and generalized additive models.

1 Introduction

Computer and simulation experiments are becoming the preferred method for analyzing systems for which physical experimentation is usually not feasible. Computer experiments are based on computer codes for which a given set of inputs generates the output(s) frequently in a deterministic manner [1, 2]. Therefore in the analysis of computer experiments, interpolation models such as Kriging are used to guarantee the zero prediction error at the data points [3, 4, 5]. In some applications however the outcome is stochastic. In stochastic simulation models for example a seed controls a random number stream and changing the seed results in different outcomes. There are also applications where the factors can be separated into two groups as "controllable" and "uncontrollable" based on their characteristics in the physical system. The uncontrollable factors could for example be the customer arrival rate in a grocery store or the room temperature in a laboratory and the controllable factors could for example be the number of checkout counters. Since the uncontrollable factors can not be controlled in the actual physical system, their input values in the simulation model have to be varied. These uncontrollable factors are different from the controllable factors and thus need to be treated differently in the analysis as well as when

designing the experiments. The analysis of the uncontrollable factors is the primary focus of this paper.

Kleijnen [3, 5] considers simulation models as a special class of computer models, which typically include one or more stochastic elements. The sources of variation are the seed controlling the random number generator and the set of uncontrollable factors that are included in the computer model to account for the environmental variations of the underlying physical system. The variation in the output from varying the seed in a simulation model originates from the embedded stochastic components such as queues, arrival processes and procedures and can be considered to correspond to the experimental error in a physical experiment. The second type of variation in simulation models is coming from changes in the uncontrollable factors. To mimic the uncertainty from the environmental factors in the physical system the settings of the uncontrollable factors are varied in the simulation model (see section 5). The variation associated with the uncontrollable factors is important for robustness [6], since the results from a simulation model generally need to be reliable under different environmental settings in the actual physical system.

The functional relationship between the uncontrollable factors and the outcome is often left unspecified and considered merely as a source of variation. Kleijnen [5] suggests for robustness analysis to summarize the mean and variance for each controllable factor settings and model them by two separate second order polynomials. Another approach is to model the functional relationship between the outcome and the uncontrollable factors. This may unveil the important uncontrollable factors. But more importantly it may unveil important interactions between controllable and uncontrollable factors, which may then be used to set the system in a more robust operating mode.

In this article the sources of variation are quantified by means of a linear mixed effects model to separate the variation into a component corresponding to changing the uncontrollable factor settings and a component corresponding to changes in the seed. Additionally, a generalized additive model is used to model the functional relationship between the outcome and the uncontrollable factors, which replaces the variance components in the linear mixed effects model.

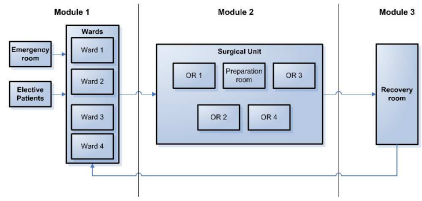


Figure 1: Basic layout of surgical unit. The patient flow is from left to right.

2 The case study

The computer model considered in this paper is a discrete event simulation model of an orthopaedic surgical unit. The model simulates the individual patient's flow through the unit (illustrated in Figure 1) and has been developed in collaboration with the medical staff at Gentofte University Hospital in Copenhagen. The unit undertakes both acute and elective (planned) surgery and performs more than 4,600 operations a year. The patients come from several wards throughout the hospital, but the main sources of incoming patients are the four orthopedic wards and the emergency care unit.

2.1 Input factors

The simulation model has several noise sources; these can be separated into noise caused by variations in the uncontrollable factors and noise caused by variation in the seed. The seed controls the random number stream embedded in the simulation model and thus variations influence the embedded queues and processes and mimic the experimental error in a physical experiment. The uncontrollable factors are for example the arrival rate of acute patients and the cleaning time of the operating rooms (ORs). Moreover, a set of controllable factors, for example the number of operating rooms and the number of surgeons, is influencing the performance of the model. The factors in the model are summarized in Table 1, which shows that the majority of the uncontrollable factors are related to resources being shared with other segments of the department and other departments of the hospital and hence might be occupied for other tasks. The outcomes from the simulation model are waiting times, patient throughput (the total number of patients treated) and the amount of overtime used on elective surgery. The simulation model is implemented in Extend [7] and controlled from a Microsoft Excel spreadsheet with a Visual Basic for applications script.

Table 1: Controllable and uncontrollable factors used in the simulation model

Controllable factors	Uncontrollable factors
Porters	Porters occupied
Elective patients	Surgeon occupied
ORs	OR cleaning time
Recovery beds	Recovery bed occupied
Cleaning teams	Cleaning teams occupied
Anesthesiologists	Anesthesiologist occupied
Operating days	Length of procedures
Acute intake	Acute arrival rate

2.2 Performance measures

As performance measures, we consider the total throughput (TT), the percentage of elective patients treated outside regular hours (EOUT) and the extent of long waiting times. Often the longest waiting times are the most important ones since from the patient's perspective they are the most bothersome [8]. The waiting time distribution for the case study is highly right-skewed with a minimum of 0 minutes, a mean of 28 minutes, a 95 % quantile of 51 minutes and a maximum of 140 minutes.

We suggest measuring the extent of long waiting times by the Conditional Value at Risk (CVaR) measure [9]. The measure originates from finance as an extension of Value at Risk (VaR) [10, 11, 12]. Both VaR and CVaR quantify a distribution of losses for example of a portfolio of assets in a single statistic. For the set of waiting times $T_x = \{t_{x1}, \ldots, t_{xN}\}$ from the x'th run, $CVaR_\alpha(T_x)$ is defined as the expected value of the α-tail distribution of T_x [10], i.e.

$$CVaR_\alpha(T_x) = \frac{\left(\frac{i_\alpha}{N} - \alpha\right) t_{xi_\alpha} + \sum_{i=i_\alpha+1}^{N} \frac{t_{xi}}{N}}{1-\alpha} \tag{1}$$

with $t_{x1} \leq t_{x2} \leq \cdots \leq t_{xN}$, i_α is the index satisfying $\frac{i_\alpha}{N} \geq \alpha > \frac{i_\alpha - 1}{N}$. t_{i_α} is the α-quantile and in economics denoted the Value at Risk (VaR). CVaR can be seen as a compromise between the average waiting time ($\alpha = 0$) and the maximum waiting time ($1 - 1/N < \alpha < 1$), where α reflects the weight put on the longest waiting times in the sample: A high α implies fewer waiting times used in the statistic and hence more weight on the longest waiting times. In the following $\alpha = 0.95$ is used so that CVaR is the average of the 5% longest waiting times.

The two other outcomes, TT and EOUT, are quality measures. They are required to

fulfill the following quality constraints

1. At least the same number of patients treated compared to the reference setting
2. The percentage of elective patients treated outside regular hours compared to the reference setting may not increase

where the reference setting corresponds to the current setting (see section 5), i.e. corresponding to the performance under the current resource allocation at the department. The requirements are constraints that ensure that a performance improvement in terms of CVaR is not obtained by treating fewer patients or generating more overtime by treating more patients outside regular hours. In this study, we focus on estimating the size of the variations in CVaR and from which sources they arise. Moreover, we want to analyze the possibility of lowering CVaR while fulfilling the quality requirements.

3 Modeling framework

As mentioned earlier, the output from the simulation model is stochastic with two types of noise coming from the uncontrollable factors and the seed controlling the random number stream. The Kriging framework often used in analysis of computer experiments is seen not to be well suited in our case, since the output is non-deterministic. There are further complications, since in our case study most controllable factors are discrete and thus interpolation is not necessarily appropriate. The presence of uncontrollable factors implies that the factors fall in two groups with different interpretations. In this study the focus is on the uncontrollable factors and we treat the controllable factor settings as a single factor. As initial model a linear model is considered

$$y(x_{c_i}, x_{e_j}, s_k) = \beta_i + \epsilon_{jk} \tag{2}$$

where β_i is the effect of controllable setting x_{c_i} and $\epsilon_{jk} \sim N(0, \sigma_\epsilon^2)$ the residual variation. x_{c_i} is the i'th controllable factor setting, x_{e_j} the j'th environmental factor setting and s_k the seed in the k'th replicate. The model has parameters for each controllable factor setting and a single error term covering the variation due to both the seed and the uncontrollable factor setting.

The linear model does not estimate the variations related to the uncontrollable factors and the seed separately. To target both types of variations explicitly a linear mixed effects model (LME) is proposed [13]. The LME is formulated such that it quantifies

the two sources of variation by estimating the variance component for each in the following model

$$y(x_{c_i}, x_{e_j}, s_k) = \beta_i + E_j + S_k \tag{3}$$

β_i is the effect of controllable setting i, $E_j \sim N(0, \sigma_E^2)$ is the random effect of the j'th uncontrollable factor setting and $S_k \sim N(0, \sigma_S^2)$ is the variation corresponding to the seed. The model is estimated by restricted maximum likelihood estimation (REML) as described in Venables and Ripley [14].

The LME model quantifies the variation corresponding to varying the settings of the uncontrollable factors in a single term. It is estimated in the variance component σ_E^2. An alternative approach is to model the functional relationship between y and each of the uncontrollable factors. These functional relationships can for example be estimated using a generalized additive model (GAM) [15]. In this modeling framework the effects of the uncontrollable factors are modeled as non-parametric smooth additive functions and the resulting model is given as

$$y(x_{c_i}, x_{e_j}, s_k) = \beta_i + \sum_{l=1}^{m} f_l(x_{e_j}^l) + S_k \tag{4}$$

with $x_{e_j}^l$ being the j'th setting for the l'th uncontrollable factor and $S_k \sim N(0, \sigma_S^2)$ the residual or seed term. f_l is a spline based smooth function with the smoothness determined by a penalty term. By estimating the functional relationship between the uncontrollable factors and the outcome, the uncontrollable factors that are needed to be tightly controlled may be identified. But more importantly interactions between controllable and uncontrollable factors may also be estimated. The estimation of the β's and the smooth functions can for example be done with the `R`-code provided by Wood [16, 17].

The fraction of runs not fulfilling the quality requirements can also be analyzed within the GAM framework. For the output y_q, $q \in \{CVaR, TT, EOUT\}$, and the quality requirement c_q, the outcome is binary, $I(y_q < c_q)$ (1 if fulfilled and 0 if not). A GAM with a binomial distribution family is considered with the linear predictor given as

$$E\left[\log\left(\frac{p}{1-p}\right)_{ij}\right] = \beta_i + \sum_{l=1}^{m} f_l(x_{e_j}^l) \tag{5}$$

where p is the fraction of runs not meeting the requirements for a given controllable factor setting.

The advantage of using the GAM framework is the employment of the smooth functions, which for example implies that a potential complex effect of an uncontrollable

factor can be easily presented graphically. Moreover, the GAM does not impose a parametric form on the functional relationship except for the spline-based functions and the additivity, which implies that the data decides the model. Another advantage is that the controllable factor effects can be corrected for the effect of the uncontrollable factors. The disadvantage of the GAM framework is the additivity assumption, which in this paper implies that only marginal effects are considered. It is possible to expand the GAM to include functions of more than one variable and interactions with e.g. controllable factors, which could potentially lead to rather complex models.

4 Example

To illustrate our modeling framework presented in section 3, we consider a simple queuing-system operating in one of two modes: M/M/1 or M/M/2 (2 servers working in parallel). The M/M/1 (M/M/2) queue consists of a single arrival process with Poisson arrivals and one (two) server(s) with exponential service times. The arrival rate, the service rate and the number of servers are denoted λ, μ and m, respectively. In the single server system the service time is defined to be approximately half as long as the service time of the servers in the two server system, which corresponds to the server utilization, $\rho = \frac{\lambda}{m\mu}$, being constant for fixed λ.

We consider the expected waiting time in the queue, W_q, as the performance parameter of the system. The expected waiting time is known to be

$$W_q = \begin{cases} \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho^2}{\lambda(1-\rho)} & m = 1 \\ \frac{\lambda^2}{\mu(4\mu^2-\lambda^2)} = \frac{2\rho^3}{\lambda(1-\rho^2)} & m = 2 \end{cases} \tag{6}$$

instead of considering μ directly, we use ρ. On log-scale the expected waiting time in the queue is given as

$$\log(W_q) = \begin{cases} -\log(\lambda) + 2\log(\rho) - \log(1-\rho) & m = 1 \\ -\log(\lambda) + \log(2) + 3\log(\rho) - \log(1+\rho) - \log(1-\rho) & m = 2 \end{cases} \tag{7}$$

The advantage of considering the expected waiting time on log-scale is that it provides a more interpretable model that separates λ from ρ. Another advantage is that it gives a more symmetric distribution of the output, which would be the argument for transforming the data if the true model were not known. In the following we set $LW_q = \log(W_q)$ for ease of notation. We treat m as a controllable factor, and λ and ρ as uncontrollable factors since it is deemed possible to control the number of servers but not the average arrival nor the service rates. The difference in waiting time for $m = 2$ vs. $m = 1$ is $LW_q(2) - LW_q(1) = \log(\rho) - \log(1+\rho) + \log(2)$.

4.1 Design

A simulation model that can operate as both a M/M/1 and a M/M/2 queue is implemented in Extend [7]. Each run of the simulation model is run for 20000 minutes where the first 10000 minutes are used as warm up period to ensure that the waiting time is stabilized. Moreover the seed controlling the random number generator is changed before each run, which makes the simulation model stochastic.

Two experimental plans are constructed; one for each setting of m. Each experimental plan consists of a uniform design with 2 factors (λ, ρ) and 100 runs. We use uniform designs since they are robust against model misspecification and do not rely on a certain model structure [18]. The uncontrollable factor region is given as the rectangle spanned by the intervals $\lambda \in [0.67, 1]$ and $\rho \in [0.48, 0.72]$ corresponding to varying the uncontrollable factors 20 % around their average values. The simulation model takes $\mu = \frac{\lambda}{m\rho}$ as input value, but the design and analysis are done for ρ. To estimate the variation related to the random seed, 5 replications are taken for each combination of m, λ and ρ, which in total gives 1000 runs.

4.2 Results

The LM, LME and GAM models defined in section 3 are used to model the LW_q values obtained from the simulation model. The parametric part of the models is given as

$$LW_q = \beta_0 + \beta_1 I(m = 2) \tag{8}$$

where $I()$ is the indicator function. ρ and λ are included in the GAM model on their original scale with a smoother for each m, yielding the following combined model

$$\begin{aligned} LW_q = &\beta_0 + \beta_1 I(m = 2) + f_1(\lambda) I(m = 1) + f_2(\lambda) I(m = 2) \\ &+ f_3(\rho) I(m = 1) + f_4(\rho) I(m = 2) \end{aligned} \tag{9}$$

where the smooth functions are expected to be $f_1(\lambda) = f_2(\lambda) = -\log(\lambda)$, $f_3(\rho) = 2\log(\rho) - \log(1 - \rho)$ and $f_4(\rho) = \log(2) + 3\log(\rho) - \log(1 + \rho) - \log(1 - \rho)$. In the LME model each combination of m, ρ and λ corresponds to one level of E_j.

Table 2 summarizes the parameters of the models for LW_q. The estimates for the LME model show that the residual variation in the LM model for LW_q mostly consists of variation caused by varying the uncontrollable factors. The residual variation in the LM-model is split into a main component corresponding to the variation related

Table 2: Summary for modeling LW_q-results from queuing system

Model	σ_E	σ_S	β_0(SD)	β_1(SD)
LM		0.48	0.11(0.02)	−0.30(0.03)
LME	0.47	0.08	0.11(0.05)	−0.30(0.07)
GAM		0.08	0.11(0.01)	−0.30(0.01)

to the uncontrollable factors and a minor component corresponding to the variation in the seed in the LME model. The residual variance in the GAM is seen to be the same as in the LME model, which indicates that the variation related to the uncontrollable factors is modeled adequately by the smooth function. Moreover, it is seen that sum of the variance component in the LME models is comparable with the total variation in the linear model.

The estimated partial effects of ρ and λ on LW_q are shown in Figure 2 with the corresponding theoretical partial effects superimposed. It can be seen that the effects of the uncontrollable factors are close to the theoretical values of the effects. For λ some minor deviations from the expected functions are seen and the two estimated curves are not perfectly parallel. The smoothed and theoretical curves are tightly superimposed, since a simple simulation model is used and the outcome is additive. The difference between the estimated effect of λ for one and two servers is however insignificant, whereas the difference for ρ is highly significant. The model explains more than 98 % of the variation in the data and the residual variation is seen to be 0.08^2 compared to the LW_q values varying from -1.82 to 1.62.

The estimates of the parameters do also coincide with the true values. Together the models provide insight on the properties of the two queuing system, by using no prior information. In the next section, we return to the case study given in section 2 and apply the proposed approach to model the CVaR waiting times.

5 Case study continued

For the case study given in section 2, the average computer time needed for simulating 6 months of operation (with one week of warm-up) is around 7 minutes. For each run the system's performance is summarized in a set of measures, e.g. the total patient throughput, the percentage of elective patients treated outside regular hours and the CVaR waiting time. Two experimental designs are considered

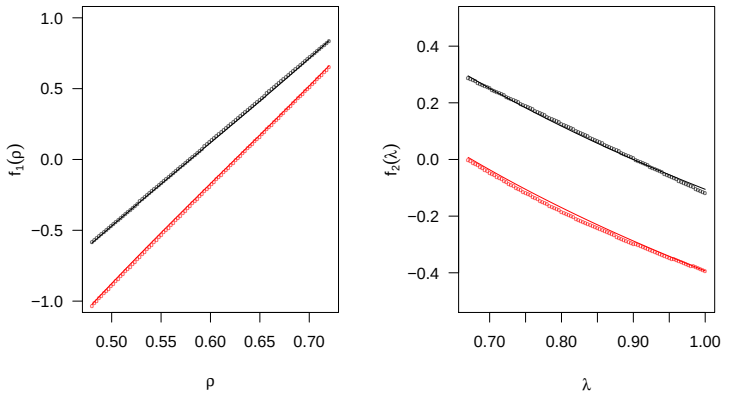


Figure 2: Estimated partial effects of ρ and λ on LW_q. Lines indicated with "o" are estimated partial effects of ρ and λ on LW_q, solid lines are the theoretical partial effects. For both ρ and λ the top curves correspond to $m = 1$ and the bottom curves to $m = 2$.

1. The current controllable factor setting corresponding to the current setup simulated with
   (a) 400 different uncontrollable factor settings chosen such that the ranges of the 8 uncontrollable factors are uniformly covered
   (b) 2 repetitions with different seeds for each setting of the uncontrollable factors, i.e. a total of $N = 800$ runs
   (c) the combined design is denoted D_C
2. 20 new controllable factor settings, which were found in a pilot study with the objective of finding good settings in terms of reducing the predicted CVaR waiting time while maintaining the performance on TT and EOUT. Each setting is simulated under
   (a) 20 different uncontrollable factor settings chosen from the 400 run uniform design with 8 factors consider in the reference design
   (b) 5 repetitions under different seeds for each uncontrollable and controllable factor combination, i.e. a total of $N = 2000$ runs
   (c) the combined design is denoted D_N

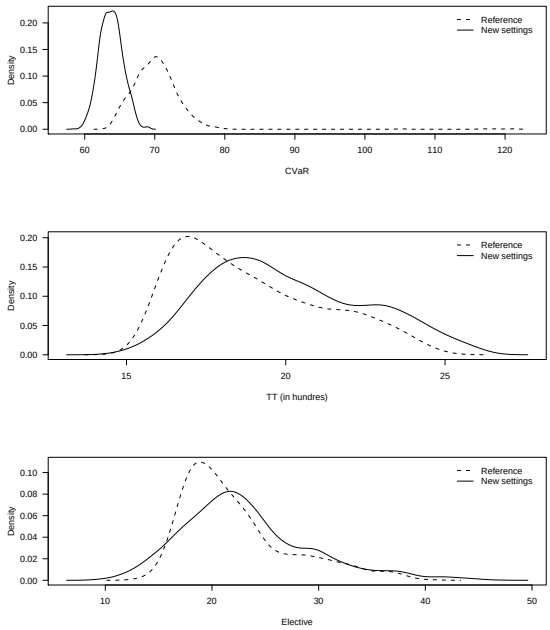


Figure 3: Estimated densities for CVaR (top), TT (middle) and EOUT (bottom) for reference design (dashed lines) and new settings (solid lines)

The sub-designs (the designs for the uncontrollable factors used for a certain setting of the controllable factor) considered in D_N are generated such that all 400 settings are assigned to one controllable factor setting each. This is done by first constructing a uniform design with 400 runs, then assigning each run to a whole plot (a combination of the settings of the controllable factors) such that all runs are assigned and each whole plot has 20 runs. The uniformity of the design is measured by the wrap-around discrepancy as suggested by Fang et al. [18]. Likewise the optimal construction of the sub-designs is achieved through the assignment of runs that minimize the maximal value of the wrap around values of the sub-designs. The main benefit of the design is that more uncontrollable factor settings can be tried compared to a crossed design, which is often used in applications with controllable and uncontrollable factors. This gives a better coverage of the uncontrollable factor space. For more detail, see Dehlendorff et al. [19].

The analysis here is focused on the output from D_N if not stated otherwise. The results from D_C serve as baseline. The outputs from both designs are shown in Figure 3 and are seen to be similar for TT and EOUT. The CVaR waiting times are seen to be lower for the new settings. Each run consists of approximately 2000 patients. Thus the CVaR waiting time becomes the average of the approximately 100 longest waiting times. The potential range for CVaR is from the 95 % quantile (51 minutes) to the maximal waiting time (140 minutes). However, as the waiting time distribution is right skewed the range of the CVaR-values goes from 55 to 80 minutes with the exception of 4 runs.

5.1 Results

Figure 3 shows the CVaR waiting times for the 20 new settings and the current settings. It can be seen that the CVaR waiting times for the new settings are shorter compared to the current setup as expected from the pilot study. Furthermore, the coefficient of variation (CV) is lower for CVaR for the new settings (CV=2.58 %) compared to the reference (CV=6.18 %). The CVs for CVaR are also seen to be lower compared to TT ($CV_{\text{cur}} = 12.01$ % and $CV_{\text{new}} = 12.23$ %) and EOUT ($CV_{\text{cur}} = 11.03$ % and $CV_{\text{new}} = 12.81$ %). For EOUT, we use the square root transformation for symmetry and consistency with Table 3. The significant increase in the CV in the reference scenario for the CVaR waiting times reflects a more right skewed distribution with observations ranging from 63.40 to 121.17 minutes. Without the 4 largest observations in the reference scenario the CV reduces to 4.13 %, which is still considerably high compared to the new settings. The overall mean CVaR is estimated to be 63.77 and 70.23 minutes with the new and reference settings, respectively. The initial analysis suggests that the new settings give lower CVaR on average and the performance is less sensitive to changes in the controllable factors.

Fitting the linear model (LM in Table 3) does indicate significant differences in mean CVaR among the 20 new settings with setting 2 having the lowest CVaR waiting time. Furthermore, the variances around the means for each setting of the controllable factors show evidence of being heterogeneous with $p = 0.005$ for Bartlett's test of variance homogeneity. From Table 3 it can also be seen that the residual variation in the reference scenario is 7 times higher compared to the new settings, which indicates that the new settings are more robust against changes in the uncontrollable factors.

Table 3: Estimate for models in section 3. The variance components are summarized in σ_E and σ_S, the overall means in μ, and the adjusted R^2 in R_a^2 for the 20 suggested settings and the reference scenario for CVaR, EOUT and TT, respectively. † Linear regression model. * The square root of the outcome is used. ‡ The linear model corresponds to the null-model since only one controllable factor setting is present.

		New settings			
Outcome	Model	σ_E	σ_S	R_a^2	μ
CVaR	LM†	-	1.63	0.02	63.77
	LME	1.17	1.16	-	
	GAM	-	1.15	0.51	
EOUT*	LM	-	0.50	0.33	4.77
	LME	0.48	0.18	-	
	GAM	-	0.18	0.91	
TT	LM	-	223.50	0.18	2005.45
	LME	224.01	42.41	-	
	GAM	-	42.74	0.97	
		Reference scenario			
		σ_E	σ_S	R_a^2	μ
CVaR	LM†	-	4.34	0‡	70.23
	LME	2.19	3.74	-	
	GAM	-	3.69	0.28	
EOUT*	LM	-	0.52	0*	4.68
	LME	0.48	0.19	-	
	GAM	-	0.18	0.88	
TT	LM	-	226.80	0*	1888.92
	LME	222.60	43.73		
	GAM	-	41.70	0.97	

5.2 LME

The REML estimates of the variance components in the LME analysis of the CVaR waiting times are also included in Table 3. The two components for CVaR are seen to be comparable in size for D_N and a bootstrapped 95 % confidence band [20] for the intraclass correlation [21] gives $0.46 \leq \frac{\sigma_E^2}{\sigma_E^2+\sigma_S^2} \leq 0.55$. For the reference setup, the variance components of the CVaR are seen to be significantly larger. This shows that not only is the current setup inferior to the proposed setups on average, but it also tends to be more sensitive to changes in the uncontrollable factors and the seed. The total reduction in variance with the new settings compared to the reference settings is 86 % with the largest relative reduction for variation corresponding to the seed being 90 %.

From Figure 3 it can be seen that the CVaR waiting times in the reference scenario have a highly right skewed distribution with 4 runs with CVaR-values greater than 100 minutes, whereas the remainder of the runs are contained in the interval [63.40, 79.87]. Furthermore, the 4 observations violate the model assumptions: $B_j \sim N(0, \sigma_E^2)$ and $S_k \sim N(0, \sigma_S^2)$. Omitting the observations from the analysis gives $\sigma_E^2 = 2.36^2$ and $\sigma_S^2 = 1.67^2$ which means an increase in σ_E^2 and a decrease in σ_S^2 with the average also decreasing from 70.23 to 70.00. The reduction in total variation without the 4 observations from the reference settings to the new settings is 68 %. The diagnostics after omitting the observations do not indicate problems with the model assumptions. It is seen that the current setting (with or without the 4 observations) is more sensitive to changes in the uncontrollable factors. The size of the variance components for TT and EOUT are equivalent for the two experiments, whereas the sample means are higher with the new settings.

5.3 GAM

To identify the significant uncontrollable factors a GAM model is fitted to the CVaR waiting times. From the estimated model it can be seen that 4 uncontrollable factors are significantly affecting the CVaR waiting times while the remaining 4 uncontrollable factors do not have an effect on the CVaR waiting times. The significant factors are the incoming rate of acute patients and the amount of time the anesthesiologists, porters and the recovery beds that are occupied by other processes.

The estimated effects of the significant uncontrollable factors are shown in Figure 4. The curves fitted individually for each of the 5 repetitions for the new settings show

that the functional form is consistent from one repetition to the next. In Table 3 it can also be seen that the residual variation is estimated to be $\sigma_S^2 = 1.15^2$, which matches the component from the LME model. This compared to the LME indicates that no information is lost by restricting the smooth functions to be additive. Moreover, the adjusted R^2's show that the benefit of including the uncontrollable factors is significant with 50 % or more improvements in R^2 compared to the linear model.

From Figure 4 it can be seen that the same functional relationships exist for both the current setting and the new settings for the occupancy of the anesthesiologists and the porters. The occupancy of the recovery beds has a steeper increase in CVaR in the reference settings compared to the new settings, which is likely to be caused by the fact that fewer beds are available in the reference settings. The smoothed curves for the occupancy of the recovery beds show that the new settings are more robust against variations in this factor. It can further be seen that the new settings are less sensitive to the arrival rate of the acute patients (Figure 4(a)). Moreover, it can also be seen from the curves for the occupancy of the porters and the anesthesiologists that the curves for the new settings are flatter compared to the current settings. This indicates an interaction between the controllable and the uncontrollable factors, and shows that with the new controllable settings the system is more robust against changes in the arrival rate and the occupancy of the recovery beds. Compared to Figure 2, the effect of increasing the arrival rate shown in Figure 4(a) corresponding to shortening the time between arrivals, is similar to the M/M/1 and M/M/2 queues for which it also increases the waiting time.

5.4 Risk profiles

The risk profiles of CVaR, TT and EOUT for each combination of the controllable factor settings are shown in Figure 5. The risks are defined as the risk of not fulfilling the quality requirements defined in section 2.2. In addition to the already defined requirements, we require that the new settings have a lower CVaR-value than the 5 % quantile in the reference setting (65.43 minutes). From Table 3 it can be seen that the performances in mean value and variance components are similar for TT and EOUT with the new settings compared to the baseline scenario. On average the TT is 6 % better (higher) in the new settings, whereas EOUT is 2 % worse (higher). This implies that it can be expected that meeting the requirement for EOUT will be more challenging.

The risks are estimated with the model in equation (5), which estimates the effect of the uncontrollable factors on the linear predictor with smooth functions. For the risks

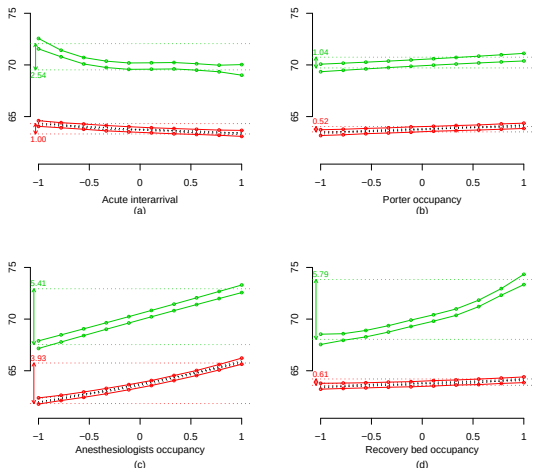


Figure 4: Estimated effects of the significant uncontrollable factors. (a) Acute inter-arrival time, (b) amount of time porters are occupied by other procedures, (c) amount of time anesthesiologists are occupied by other procedures and (d) amount of time the recovery beds are used for other patients. The two top curves in each of the 4 sub figures correspond to the 95 % confidence limits for the estimated effects in the reference design. The bottom curves consist of two solid curves corresponding to a 95 % confidence limits for the estimated effect in a model with all 5 repetitions in D_N included and 5 dashed curves corresponding to a model for each of the 5 repetitions.

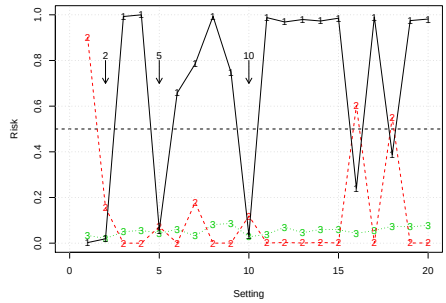


Figure 5: Risk profiles for CVaR (dotted line marked *3*), TT (dashed line marked *2*) and EOUT (solid line marked *1*). Arrows indicate settings with risks lower than 0.5 (marked by dashed line) for TT and EOUT.

corresponding to TT and EOUT, it can be seen that settings 2, 5 and 10 perform well for both measures. It can also be seen that the TT and EOUT risks are negatively correlated with Spearman's rho is equal to -0.89, i.e. that lowering the risk of treating too few patients increases the risk of treating more elective patients outside regular hours. Settings 2, 5 and 10 are quite similar, that is they operate with 4 operating days, 4 operating rooms and an increase in elective patients by 2 per day. The 3 settings use more resources compared to the current setup with the lowest additional costs for setting 5. It can be seen that settings 2, 5 and 10 on average fulfill all the requirements in more than 80 % of the runs. Compared to the reference setting the most interesting difference in the controllable factors is the use of 4 operating days compared to 5 as in the current setting.

6 Conclusion

In this article, we present the analysis of a simulation model with two types of variation due to changing seed and changes in the settings of the uncontrollable factors. The usefulness of using a generalized additive model and a linear mixed model models were illustrated by a theoretical queuing system, which showed that the suggested modeling framework performed equally well for the well-known queuing systems. The analysis for our case study shows that the variation in the CVaR waiting time with a linear mixed effects model can be split into two equally large variance components

for a set of new settings, whereas the seed variance in the reference scenario is lower compared to the variance caused by changes in uncontrollable factors. A generalized additive model shows that the main source of variation for the new settings is the use of the anesthesiologist for other tasks. Moreover, the new settings eliminate the impact of one of the most important uncontrollable factors.

The use of the linear mixed effects model provides additional insight on the variation related to the settings of the uncontrollable factors and the generalized additive model identifies the most important uncontrollable factors. This may assist decision makers in constructing focused strategies for controlling the uncontrollable factors better and if possible to improve the robustness of the system. In this application for example to ensure a more reliable access to the anesthesiologist seemed to be beneficial. The analysis also shows that the uncontrollable factors interacted with the controllable factors. Given the new settings the system was deemed more robust to changes in the uncontrollable factors.

Moreover, specific settings of the controllable factors improved the long waiting times significantly while keeping a low risk of treating fewer patients or more patients outside regular hours. The drawback of the improvements in the CVaR waiting time was the cost of the additional resources needed. By combining cost and performance, it may be possible to find cost-effective solutions balancing cost and waiting time. The cost-effectiveness issue is important for further analysis as resources are a constraint. This could be done by translating waiting time into cost or by letting waiting time serve as a risk measure in a Pareto frontier analysis.

References

[1] Santner TJ, Williams BJ, Notz WI. *The Design and Analysis of Computer Experiments*. Springer, 2003.

[2] Sacks J, Welch WJ, Mitchell TJ, Wynn HP. Design and analysis of computer experiments. *Statistical Science* 1989; **4**(4):409–423.

[3] Kleijnen JP. Kriging metamodeling in simulation: A review. *European Journal of Operational Research* 2009; **192**(3):707–716.

[4] Martin JD, Simpson TW. Use of kriging models to approximate deterministic computer models. *AIAA Journal* 2005; **43**(4):853–863.

[5] Kleijnen JP. *Design and Analysis of Simulation Experiments*. Springer, 2008.

[6] Sanchez SM. Robust design: Seeking the best of all possible worlds. In *Proceedings of the 2000 Winter Simulation Conference*. 69–76.

[7] Krahl D. The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*. 205–213.

[8] Bielen F, Demoulin N. Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality* 2007; **17**(2):174–193.

[9] Dehlendorff C, Kulahci M, Merser S, Andersen KK. Conditional value at risk as a measure for waiting time in simulations of hospital units. *Quality Technology and Quantitative Management* 2009; Submitted, in review.

[10] Rockafellar RT, Uryasev S. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance* 2002; **26**:1443–1471.

[11] Kibzun A, Kuznetsov E. Comparison of var and cvar criteria. *Automation and Remote Control* 2003; **64**(7):153–164.

[12] Kibzun AI, Kuznetsov EA. Analysis of criteria var and cvar. *Journal of Banking & Finance* 2006; **30**(2):779–796.

[13] Pineiro J, Bates D. *Mixed Effects Models in S and S-PLUS*. Springer, 2000.

[14] Venables WN, Ripley BD. *Modern Applied Statistics with S*. Springer-Verlag, 2002.

[15] Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall/CRC, 1990.

[16] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[17] Wood S. *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC, 2006.

[18] Fang KT, Li R, Sudjianto A. *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC, 2006.

[19] Dehlendorff C, Kulahci M, Andersen KK. Designing simulation experiments with controllable and uncontrollable factors. In *Proceedings of the 2008 Winter Simulation Conference*.

[20] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman & Hall, 1993.

[21] Montgomery DC. *Design and Analysis of Experiments.* 7th edition. John Wiley and Sons, Inc, 2009.

Paper F

2-stage approach for Kriging for simulation experiments with quantitative and qualitative factors

---

Working paper

2-stage approach for Kriging for simulation experiments with quantitative and qualitative factors

Christian Dehlendorff Murat Kulahci

Klaus K. Andersen

Abstract

Kriging is often used to obtain meta-models for deterministic simulation. In this article we propose a procedure that handles simulation experiments with both quantitative and qualitative factors, i.e., with the input domain divided into two strata. The proposed procedure relies on the usual Kriging framework, but introduces an initial step to assess the similarity of the model segments, which is used in the estimation of a combined model over all segments.

key words: Computer experiments, kriging, meta-modeling, simulation model

1 Introduction

Computer experiments have been receiving increasingly more attention with the growing use of computationally expensive computer models to simulate complex systems (Sacks et al., 1989; Santner et al., 2003; Martin and Simpson, 2005). Often these expensive computer models are replaced by cheaper meta-models, which are better suited for analysis and optimization. Computer experiments are often assumed to give deterministic output, which implies that a natural criterion for the meta-models is to interpolate the data. A method originating from geo-statistics called Kriging, developed by Krige and improved by Matheron (1963), is often applied in the field of computer experiments (Martin and Simpson, 2005; Sacks et al., 1989; Santner et al., 2003). The usual Kriging model is an interpolator and can fit complex responses surfaces, which makes it a model well suited for deterministic computer experiments.

Simulation models are a subtype of computer models, which can be analyzed within the Kriging framework (Kleijnen (2008a,b, 2009); van Beers and Kleijnen (2008); Ankenman et al. (2008) and Johnson et al. (2008)). Simulation models are usually divided into two subcategories; deterministic and stochastic. In deterministic simulation the output is observed without uncertainty and hence interpolation is a desired property, whereas in stochastic simulation replicates give different outputs and therefore the objective is to fit a predictor for the underlying signal. The variation in the output in stochastic simulation is caused by stochastic components such as arrival processes and queues. Stochastic simulation models are analyzed by for example Kleijnen

(2008a) and Ankenman et al. (2008). The former uses the usual Kriging framework on the averages at each design site and bootstraps to estimate the true predictor variance, whereas the latter expand the Kriging model with an extra term corresponding to the replication variation. In this paper, we only consider deterministic output by means of a discrete event simulation model for an orthopedic surgical unit at a hospital (Dehlendorff et al., 2010b) given in section 6.

A subtype of simulation models with two factor types; qualitative and quantitative is considered in this paper. This is not handled in the usual Kriging framework, which assumes that all factors are quantitative. Moreover, the response surface may be different from one level of a qualitative factor to the next, which implies that unrestricted interpolation across the levels of the qualitative factors may not be appropriate. On the other hand some correlation is expected between the levels of the qualitative factors and hence treating these levels independently is not appropriate either. In this article a novel method, which uses methods from the usual Kriging framework in a two stage estimation method for experiments with two types of input factors, is proposed.

Hung et al. (2009) and Qian et al. (2008) consider another framework for Kriging for computer models with qualitative and quantitative factors. They use the levels of the qualitative factors to define the closeness of the observations together with the usual correlation function for the continuous factors. Hung et al. (2009) focus on computer experiments with branching and nested factors, where the branching factors can be seen as a special case of having

qualitative factors. A different approach for modeling computer models with quantitative and qualitative factors is given by Zhou et al. (2010). They use a penalty based on a hypersphere parameterization. We discuss this method in detail in section 4.2.

We start by introducing the case-study in section 2 and the usual Kriging model in section 3. In section 4 the Kriging framework is expanded to handle quantitative and qualitative factors. The new framework is compared to the methods suggested by Hung et al. (2009) (Qian et al. (2008)) and Zhou et al. (2010) on a set of test functions in section 5 and on a specific application in section 6. with results indicating that our method gives more accurate meta-models.

2 Case-study

In this section we consider a discrete event simulation model for an orthopedic surgical unit at a hospital. The basic outline of the surgical unit is illustrated in Figure 1 and consists of three main modules: arrival, operating facilities and recovery.

The model simulates the patient route through the unit and the model consists of eight qualitative factors such as the staffing, the number of operating rooms and recovery beds (we treat these factors as qualitative, since only a few levels are present for each factor) and eight quantitative factors such as the incoming rate of acute patients (the factors are given in Table 1). The eight quantitative factors are uncontrollable in the physical system and hence

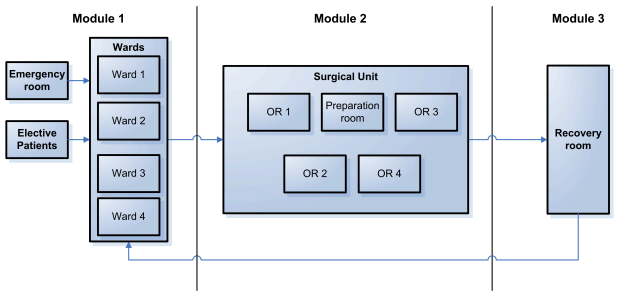


Figure 1: Surgical unit

the system can only be controlled only through the eight qualitative factors, e.g., making the system robust is done by setting the qualitative factors (see for example (Dehlendorff et al., 2010a, 2011)). In this article we however only deal with the qualitative/quantitative aspect of the model and for robustness issues we refer to Dellino et al. (2009).

Type	Factors	
Controllable	*Porters*	*Anesthesiologists*
	ORs	*Recovery beds*
	Cleaning teams	*Elective patients*
	Operating days	*Acute intake*
Uncontrollable	*Porters occupied*	*Anesthesiologist occupied*
	OR cleaning time	*Recovery bed occupied*
	Cleaning teams occupied	*Surgeon occupied*
	Length of procedures	*Acute arrival rate*

Table 1: Factors used in simulation model for surgical unit

In this simulation study the primary concern is the extent of long waiting times, which is measured by the Conditional Value of Risk (CVaR) waiting time as described in Dehlendorff et al. (2010b). The measure is a statistic used in finance for example to quantify a distribution of losses in portfolio optimization (Kibzun and Kuznetsov, 2003, 2006; Alexander et al., 2006).

The measure corresponds to the sample average of the 5 % longest waiting times and is a compromise between using the overall sample average (called a risk neutral strategy) and the sample maximum (called a risk averse strategy).

The simulation model is kept in a deterministic operating model by keeping the seed controlling the random number generator fixed. A single run corresponds to approximately 2000 surgical procedures and takes around seven minutes to complete, which implies that trying all possible settings is simply computationally unfeasible. The model is implemented in Extend (Krahl, 2002) and controlled from an Excel spreadsheet by a Visual Basic for Applications script.

3 Kriging

In this section we briefly introduce Kriging (for further details see Sacks et al. (1989); Kleijnen (2008a) and Santner et al. (2003)). Kriging is a modeling method that approximate a deterministic function (model) with a random function (Santner et al., 2003), but for practical reasons we will use Kriging as the acronym for the modeling framework. We estimate the model with the Matlab toolbox DACE (Lophaven et al., 2002a,b), which is one of the commonly used publicly available toolboxes for Kriging.

We consider a function or computer code that, given the input vector $\mathbf{x}$, generates the scalar and deterministic output $y(\mathbf{x})$. The Kriging model relies on the assumption that the deterministic output $y(\mathbf{x})$ can be described by

the random function

$$Y(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\boldsymbol{\beta} + Z(\mathbf{x}) \tag{1}$$

where $\mathbf{f}(\mathbf{x})^T\boldsymbol{\beta}$ is a parametric trend with p parameters and $Z(\mathbf{x})$ is a zero mean gaussian random field assumed to be second order stationary with covariance function $\sigma^2 R(\mathbf{x_i}, \mathbf{x_j})$ (Santner et al., 2003; Ankenman et al., 2008). We will return to the correlation structure in section 4.2. $Y(\mathbf{x})$ is a random field required to interpolate the true function at the design sites. The interpolation property is one of the main advantages of using Kriging for deterministic computer models.

We consider a set of n design points $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ and corresponding observations $\mathbf{y} = \{y(\mathbf{x_1}), \ldots, y(\mathbf{x_n})\}$ where $y()$ is the true function (computer model). The correlation matrix for the design points is denoted $\mathbf{R}(\boldsymbol{\theta})$ where the (ij)th element is the correlation between the ith and jth design points given as $R(\mathbf{x_i}, \mathbf{x_j})$. Likewise the vector of correlations between the point, $\mathbf{x}$, and the design points is defined as

$$\mathbf{r}(\mathbf{x}) = [R(\mathbf{x_1}, \mathbf{x}), \ldots, R(\mathbf{x_n}, \mathbf{x})]^T \tag{2}$$

The regressor $\mathbf{f}(\mathbf{x})$ is given by a vector with p regressor functions $[f_1(\mathbf{x}) \ldots f_p(\mathbf{x})]^T$ and the regressors for the design sites are given by $\mathbf{F} = [\mathbf{f}(\mathbf{x_1})^T \cdots \mathbf{f}(\mathbf{x_n})^T]^T$. Usually ordinary Kriging is used and hence $\mathbf{f}(\mathbf{x})$ reduces to $f(x) = 1$ corresponding to the model $Y(\mathbf{x}) = \mu + Z(\mathbf{x})$.

The correlation function is parameterized by a set of parameters $\boldsymbol{\theta}$, which is described in more detail in section 4.2. Given $\boldsymbol{\theta}$, the restricted maximum like-

lihood estimate of $\boldsymbol{\beta}$ (Santner et al., 2003) (assuming a gaussian distribution) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\hat{\mathbf{R}}(\boldsymbol{\theta})^{-1}\mathbf{F})^{-1}\mathbf{F}^T\hat{\mathbf{R}}(\boldsymbol{\theta})^{-1}\mathbf{y} \tag{3}$$

where $\hat{\mathbf{R}}(\boldsymbol{\theta})$ is the correlation matrix for the design sites and parameterized by the estimated parameter vector $\boldsymbol{\theta}$. The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}})^T\hat{\mathbf{R}}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) \tag{4}$$

where n is the number of observations and p is the rank of F (the number of parameters in $\hat{\boldsymbol{\beta}}$). $\hat{\sigma}^2$ is seen to be adjusted for the number of parameters in the parametric part of the model. The correlation parameters are found by minimizing the negative restricted profile log-likelihood (L_r) for $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[(n-p)\log\hat{\sigma}^2 + \log(|\mathbf{R}(\boldsymbol{\theta})|)\right] \tag{5}$$

where $|\mathbf{R}(\boldsymbol{\theta})|$ is the determinant of the correlation matrix corresponding to the design points. Given $\hat{\mathbf{R}}(\boldsymbol{\theta})$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ the predictor at $\mathbf{x}$ is

$$\hat{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x})^T\hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{x})^T\hat{\mathbf{R}}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mathbf{F}\hat{\boldsymbol{\beta}}) \tag{6}$$

At a design point $\mathbf{x} \in \mathbf{X}$ the vector $\mathbf{r}(\mathbf{x})^T\hat{\mathbf{R}}(\boldsymbol{\theta})^{-1}$ consists of $(n-1)$ zeroes and a single one at the index corresponding to $\mathbf{x}$, which implies that the predictor is $y(\mathbf{x})$.

4 Kriging with qualitative and quantitative factors

In this section, we consider Kriging for computer models with qualitative factors (or at least ordinal factors with few levels) and quantitative factors. This is often the case for simulation models, e.g., the number of operating rooms at a surgical unit at a hospital vs. the incoming rate of acute patients to the unit. The output from such a model depends on both qualitative and quantitative factors. Even though the simulation may behave differently from one combination of the qualitative factors to another, some correlation between observations having different qualitative factor settings is expected. The setup is similar to a split-plot experiment in which a combination of the qualitative factors corresponds to a whole plot and a combination of the quantitative factors is a subplot.

We now consider a set of observations of size $n = mq$ with m qualitative factor combinations and q quantitative factor settings. In this setup, for a given combination of settings for the qualitative factors (a whole plot), experiments are run at various settings of the quantitative factors resulting in n different quantitative factor settings in the combined design. For a more detailed explanation of such a set up, see Dehlendorff et al. (2008, 2011). To ease the notation in the following, we will denote a combination of the qualitative factors a "whole plot", but note that the experimental design is not a split-plot design. We furthermore assume that the observations are ordered by whole plot. Hence the input consists of two components, where

w_i is the whole plot or qualitative component and x_{ij} the quantitative part.

4.1 Model

For a model with qualitative and quantitative factors, we assume that the Kriging predictor of interest is of the form

$$\hat{y}(\mathbf{w}_i, \mathbf{x}_{ij}) = \mathbf{f}(\mathbf{w}_i)\hat{\boldsymbol{\beta}} + \mathbf{r}(\mathbf{w_i}, \mathbf{x_{ij}})^T \hat{\mathbf{R}}(\boldsymbol{\theta})^{-1}(\mathbf{y_x} - \mathbf{F}\hat{\boldsymbol{\beta}}) \tag{7}$$

where $f(\mathbf{w}_i)$ depends purely on the whole plot setting. Here the special case $\mathbf{f}(w_i) = [1 \quad I(w_i = 2) \quad \cdots \quad I(w_i = m)]$ is considered in which $I()$ is the indicator function and w_i the whole plot number of observation i (a scalar $w_i \in \{1, \ldots, m\}$). $\boldsymbol{\beta}$ consists of $[\mu_1, \tau_2, \ldots, \tau_m]$, where μ_1 is the expected value for whole plot 1 and $\mu_2 = \mu_1 + \tau_2$ the expected value for whole plot 2, etc. The parametric structure is introduced to handle the difference in the output from one whole plot to the next, but without assuming a structure for the qualitative factors. To simplify the notation in the remainder of the paper we denote the jth quantitative factor settings (the quantitative factor settings in the jth subplot) in the ith whole plot $\mathbf{x}_{ij}$. Moreover, the input matrix $\mathbf{X}$ is a matrix consisting of the quantitative component of the input ordered by whole plot

$$\mathbf{X} = \begin{bmatrix} \mathbf{x_{11}}^T & \mathbf{x_{12}}^T & \cdots & \mathbf{x_{1q}}^T & \mathbf{x_{21}}^T & \cdots & \mathbf{x_{m(q-1)}}^T & \mathbf{x_{mq}}^T \end{bmatrix}^T \tag{8}$$

that is; $\mathbf{X}$ is a $(mq) \times d_x$ matrix, where d_x is the number of quantitative factors.

4.2 Correlation structure

For a simulation experiment with m whole plots (i.e., qualitative factor combinations) and q quantitative factor combinations within each whole plot (having the same number of quantitative factor combination is not a requirement for the method but it eases the notation in the following), we now address how the correlation between two observations from different whole plots could be defined. First, we consider the simple situation with two observations from the same whole plot: $\mathbf{x}_{ij}$ and $\mathbf{x}_{il}$. If the simple Gaussian correlation correlation structure is used the correlation between two observations within the same whole plot is given as

$$\tilde{R}(\mathbf{x}_{ij}, \mathbf{x}_{il}) = \exp\left(-\sum_{p=1}^{d_x} \theta_p (x_{ij}^p - x_{il}^p)^2\right) \tag{9}$$

where d_x is the number of quantitative factors and θ_p is the correlation parameter for the pth quantitative factor (see for example Sacks et al., 1989).

Observations from different whole plots are not expected to be as correlated as observations coming from the same whole plot. This implies that the correlation in Equation (9) should be reduced by a factor depending on the similarity of the qualitative factor settings

$$R(\mathbf{x}_{ij}, \mathbf{x}_{kl}) = \tilde{R}(\mathbf{x}_{ij}, \mathbf{x}_{kl}) \cdot (I(i = k) + I(i \neq k)\alpha_{ik}) \tag{10}$$

where $\tilde{R}(\mathbf{x}_{ij}, \mathbf{x}_{kl})$ is the correlation function in equation (9) evaluated as if the observations were from the same whole plot, $I()$ is the indicator function and $0 \leq \alpha_{ik} \leq 1$. Three simple ways of defining α_{ik} are

1. $\alpha_{ik} = 0$: x_{ij} and x_{kl} are uncorrelated for $i \neq k$

2. $\alpha_{ik} = \theta_c$: same correlation reduction for observations from different whole plots, where $\theta_c \in [0, 1]$

3. $\alpha_{ik} = 1$: no reduction

Clearly correlation structures 1 and 3 are special cases of correlation structure 2 and hence we only need to consider structure 2. In correlation structure 2 the θ_c-parameter is estimated together with the other correlation parameters. One issue in the choice of α_{ik} is that the resulting correlation matrix should be positive definite (Qian et al., 2008), which is ensured by the correlation structure in (10).

Hung et al. (2009) (HRM) propose a different correlation function, which is developed for computer experiments with branching, nested and shared factors. Of theses factors the branching factors are considered to be qualitative factors in this study. If one disregard the nested factor aspect the computer model in this study can be analyzed using their model. HRM propose the following correlation function for the Kriging model

$$R((z_i, x_i), (z_k, x_k)) = \exp\left(-\sum_{p=1}^{d_x} \theta_p (x_i^p - x_k^p)^2\right) \exp\left(-\sum_{q=1}^{d_z} \theta_{zq} I(z_i^q \neq z_k^q)\right) \tag{11}$$

where z_i^q is the qth qualitative/branching factor and x_i^p the pth quantitative/shared factor for observation i and $I()$ is the indicator function. With one qualitative factor this is seen to be similar to the correlation structure with $\alpha_{ik} = \theta_c$.

Zhou et al. (2010) (ZQZ) consider a hypersphere parameterization of the correlation between observations with different qualitative factor levels. They consider the combinations of the qualitative factor levels as a single categorical variable with m levels. The correlation structure has the same structure as in equation (10), where α_{ik} is given by the (ik)th element of matrix $\mathbf{T}$. The penalty matrix is constructed by the hypersphere decomposition in two steps. Step 1 is a Cholesky decomposition $\mathbf{T} = \mathbf{L}\mathbf{L}^{\mathbf{T}}$ and step 2 is the construction of the lower triangular matrix $\mathbf{L}$ given as

$$L_{rs} = \begin{cases} 1 & r = s = 1 \\ \cos(\theta_{r,s}) & s = 1 \ (r > 1) \\ \sin(\theta_{r,1}) \cdots \sin(\theta_{r,s-1}) \cos(\theta_{r,s}) & s = 2, \ldots, r-1 \ (r > 1) \\ \sin(\theta_{r,1}) \cdots \sin(\theta_{r,r-2}) \sin(\theta_{r,r-1}) & r = s \ (r > 1) \end{cases} \tag{12}$$

where L_{rs} is the (rs)th element of $\mathbf{L}$ and $\theta_{r,s} \in [0, \pi]$. This way $\mathbf{T}$ is ensured to be positive definite matrix with unit diagonal elements and hence the correlation function in equation (10) is a valid correlation function. The correlation structure can handle both negative and positive correlations between observations from different levels of the categorical factors. One drawback of the method is the number of correlation parameters needed for $\mathbf{T}$ is given as $1/2m^2 + 1/2m - 1$, e.g., 209 parameters are required to be estimated for 20 qualitative factor settings. This implies that the model requires a lot of data and estimation may become slow.

A simpler approach is to use the sample averages and standard deviations for

each whole plot as a measure of their similarity. This implies that α_{ik} may be defined as $\alpha_{ik} = \exp\left(-\theta_{\hat{\mu}}(\hat{\mu}_i - \hat{\mu}_k)^2 - \theta_{\hat{\sigma}}(\log(\hat{\sigma}_i) - \log(\hat{\sigma}_k))^2\right)$, where $\hat{\mu}_i$ and $\hat{\sigma}_i$ are the sample average and standard deviation for the ith whole plot (log-transformed to make it robust to outliers). This correlation structure is motivated by the fact, that we expect similar whole plots to have the similar average and standard deviations, i.e., observations with similar mean and standard deviation are also expected to be correlated.

The mean-standard deviation model can be estimated within the usual Kriging framework by augmenting the input matrix $\mathbf{X}$ in (8) with a matrix $\mathbf{M}$

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \mathbf{M} \end{bmatrix} \tag{13}$$

where

$$\mathbf{M} = \begin{bmatrix} \hat{\mu}_1 & \log(\hat{\sigma}_1) \\ \hat{\mu}_2 & \log(\hat{\sigma}_2) \\ \vdots & \vdots \\ \hat{\mu}_m & \log(\hat{\sigma}_m) \end{bmatrix} \otimes \mathbf{1}_{q \times 1} \tag{14}$$

and fit the Kriging model on $\tilde{\mathbf{X}}$. It can be seen that the model allows for predictions for whole plots not already observed provided that estimates for the mean and standard deviation are available. This can be handled by the ZQZ-model, but requires correlation parameters for the correlation between the new whole plot and all existing whole plots, which may be more difficult to give.

4.3 2-stage procedure

Instead of using the average and standard deviations as whole plot similarity measures as suggested in section 4.2, one could argue that the similarity between the whole plots should be judged on a measure relating to the correlation structure. Instead of using the levels of the whole plot factors or the sample mean and standard deviation, the similarity of observations from the different whole plots is measured by the similarity of the correlation function parameters for the whole plots. This can be done with a procedure in two stages: 1) fit m Kriging models for the quantitative factors in the m subsets of the data corresponding to m whole plots and 2) use the correlation parameters estimated in these m Kriging models as similarity measures. The first stage gives m models for the quantitative factors in each whole plot

$$Y_i(\mathbf{x}_{ij}) = \mu_i + Z_i(\mathbf{x}_{ij}) \quad i = 1, \ldots, m \tag{15}$$

where $Z_i()$ has the correlation function

$$R_i(\mathbf{x}_{ij}, \mathbf{x}_{ik}) = \exp\left(- \sum_{p=1}^{d_x} \theta_{ip} (\mathbf{x}_{ij}^p - \mathbf{x}_{ik}^p)^2 \right) \quad i = 1, \ldots, m \tag{16}$$

This gives a matrix of correlation parameters

$$\boldsymbol{C} = \begin{bmatrix} \theta_{11} & \cdots & \theta_{1d_x} \\ \vdots & \ddots & \vdots \\ \theta_{m1} & \cdots & \theta_{md_x} \end{bmatrix} \tag{17}$$

where θ_{ij} is the correlation parameter for the jth quantitative factor in the ith whole plot and $\mathbf{C}_i$ the correlation parameters for the ith whole plot. The intuition is that similar whole plots tend to have similar correlation parameters and thus the difference in the correlation parameters determines the correlation. To measure the whole plot similarity the information in the $\mathbf{C}$-matrix is added to the original design sites $\mathbf{X}$ such that the design sites are given as

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{C}} \end{bmatrix} \tag{18}$$

where

$$\tilde{\mathbf{C}} = \mathbf{C} \otimes \mathbf{1}_{q \times 1} \tag{19}$$

This can straightforward be generalized to the general case where the number of quantitative factor settings tried at the whole plots is not the same for all whole plots. The combined Kriging model with $\tilde{\mathbf{X}}$ becomes

$$Y(\tilde{\mathbf{x}}_{ij}) = \mu + Z(\tilde{\mathbf{x}}_{ij}) \tag{20}$$

where $Z()$ has correlation function

$$R(\tilde{\mathbf{x}}_{ij}, \tilde{\mathbf{x}}_{kl}) = \exp\left(- \sum_{p=1}^{2 \times d_x} \tilde{\theta}_p (\tilde{\mathbf{x}}_{ij}^p - \tilde{\mathbf{x}}_{kl}^p)^2 \right) \tag{21}$$

Estimating the parameters for the models in equations (15) and (20) can be done with the methods described in section 3.

The main challenge of this method is to get reliable correlation parameters in the first stage in which the models are based on few data points. Moreover,

the time for fitting the models is an issue, since m models need to be fitted. This may however not be a problem if the number of whole plots is not too small, since the execution time of the fitting procedure is proportional to n^3 (Lophaven et al., 2002a); that is, fitting m model with n/m observations each gives an execution time in the order of n^3/m^2. In the final model the full data set is used, but from the $\boldsymbol{C}$-matrix in equation (17) a good initial guess for the d_x first correlation parameters can be found to speed up the convergence, e.g., by using the column-wise averages.

A potential benefit of using this correlation function compared to the one proposed by HRM is that it uses the correlation structure as the similarity measure instead of the levels of the qualitative factors. The latter may run into problems if the similarity of the whole plots depends for example on an interaction between two factors. Compared to the method proposed by ZQZ fewer correlation parameters are used, i.e., for m whole plots and dimension d_x, the 2-stage model uses d_x parameters to parameterize the whole plot correlation in the final model, whereas ZQZ use $m^2/2 + m/2 - 1$ parameters. Figure 2 illustrates the difference in the number of parameters needed to parameterize the whole plot correlation, which shows that for example with $m = 10$ whole plots the number of quantitative factors must be more than 54 to favor the ZQZ parameterization. The 2-stage model is considerably easier to fit compared to the model by ZQZ, but it can not handle the negative correlations between whole plots as in ZQZ. Furthermore, the ZQZ is a simpler model if the number of whole plots is limited and the number of quantitative factors is large (see Figure 2).

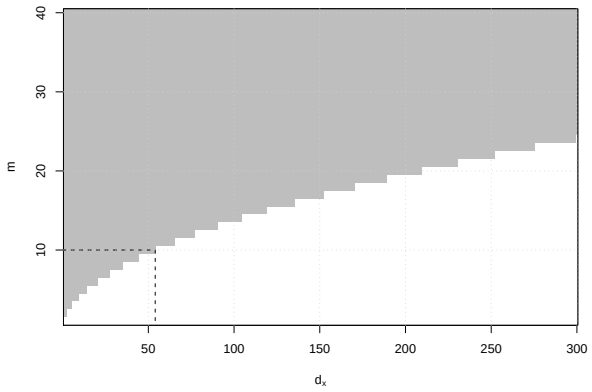


Figure 2: Comparison of correlation parameters needed for parameterizing the whole plot correlation in the ZQZ and 2-stage models. The dark area corresponds to cases in which the 2-stage model has fewer parameters

5 Test functions

In this section we consider three functions as test cases, which are listed in Table 2. They represent three situations: identical whole plots, whole plots with one active factor in common and whole plots with completely different active factors. All three cases consist of two groups of whole plots, such that whole plots from different groups are different, whereas whole plots from the same group are similar. The constant h in the sinusoidal function determines the variance of this whole plot group.

	Function		
Whole plots	1	2	3
1, 2	$x_{i1}\exp\left(-x_{i1}^2 - x_{i2}^2\right)$	$h\sin(x_{i1})$	$h\sin(x_{i3})$
3, 4	$x_{i1}\exp\left(-x_{i1}^2 - x_{i2}^2\right)$	$x_{i1}\exp\left(-x_{i1}^2 - x_{i2}^2\right)$	$x_{i1}\exp\left(-x_{i1}^2 - x_{i2}^2\right)$

Table 2: Test functions

In Table 3 the performance for four different correlation structures are compared in terms of their mean squared prediction error. Each model is based on the same training data, which has 50 observations in each whole plot. Likewise the same validation data set is used for all combinations of functions and correlation structures and consists of 10.000 randomly chosen points.

Case	Model	Function 1	Function 2	Function 3
$h = 0.56$	2-stage	$1.05 \cdot 10^{-8}$	$2.21 \cdot 10^{-4}$	$4.29 \cdot 10^{-3}$
	$\alpha_{ik} = \theta_c$	$\mathbf{8.39 \cdot 10^{-9}}$	$5.11 \cdot 10^{-4}$	$8.16 \cdot 10^{-3}$
	$\alpha_{ik} = g(\hat{\mu}_i, \hat{\sigma}_i)$	$1.12 \cdot 10^{-8}$	$\mathbf{5.04 \cdot 10^{-5}}$	$\mathbf{3.41 \cdot 10^{-3}}$
	ZQZ	$1.83 \cdot 10^{-8}$	$3.12 \cdot 10^{-4}$	$3.71 \cdot 10^{-3}$
$h = 0.15$	2-stage	$1.05 \cdot 10^{-8}$	$4.27 \cdot 10^{-5}$	$\mathbf{8.98 \cdot 10^{-4}}$
	$\alpha_{ik} = \theta_c$	$\mathbf{8.39 \cdot 10^{-9}}$	$1.48 \cdot 10^{-4}$	$2.67 \cdot 10^{-3}$
	$\alpha_{ik} = g(\hat{\mu}_i, \hat{\sigma}_i)$	$1.12 \cdot 10^{-8}$	$\mathbf{3.81 \cdot 10^{-5}}$	$1.66 \cdot 10^{-3}$
	ZQZ	$1.97 \cdot 10^{-8}$	$2.13 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$

Table 3: MSPE for test functions. $h = 0.56$ corresponds to 14 times higher variance in sinusoidal group and $h = 0.15$ to equal variance

In the first example in Table 3, the whole plot groups are designed such that the variance in the sinusoidal part of functions 2 and 3 is approximately 14 times higher than the other group (with $h = 0.56$). This should favor the mean-standard deviation correlation structure, since it uses the standard deviation in the correlation among whole plots. The results show that the correlation structure that bases the similarity of whole plots on the sample averages and standard deviations performs better than the 2-stage approach. The 2-stage procedure uses the first correlation parameter estimates to differentiate between whole plots, whereas the correlation structure using the mean and standard deviations utilizes the information contained in the difference in standard deviations.

The HRM-model corresponds to a constant reduction, since only one branching (qualitative) factor is present. In section 6 a more general model is considered for which the whole plots are generated from several qualitative factors, i.e., better suited for model considered by HRM. Finally it can be seen that the model proposed by ZQZ has a performance comparable with the 2-stage model. Figure 3 compare the correlation between whole plots estimated with the four methods. It can be seen that they are similar except for the correlation structure with a constant reduction.

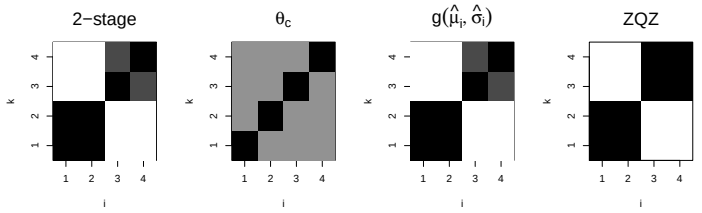


Figure 3: Correlation matrices for the correlation between whole plots corresponding to α_{ik} in equation (10) for Function 3 with $h = 0.56$. The color scale is goes from white ($\alpha_{ik} = 0$) to black ($\alpha_{ik} = 1$). In the model by ZQZ the small negative correlations (in the order of 10^{-2}) have been truncated to 0.

In the lower half of Table 3 the performances of the different correlation structures are shown for the three test functions, in which the variances of the whole plot groups are designed to be equal. It can be seen that the 2-stage method performs better in terms of MSPE compared to the other correlation structures for the third function. For the first function the all three models give the same Kriging model and the same prediction error (with some minor numerical variation). It can be seen that using the sample means and standard deviations is a viable option as long as the whole plots

are not too different. Thus it performs a little better than the 2-stage model for the second model.

In section 6, we evaluate the correlation structures on a discrete event simulation model, which illustrates the benefits of using the 2-stage Kriging model in a more realistic setting.

6 Case-study continued

We now return to the case-study from section 2 for which two experiments are considered. In the first example the whole plots are expected to be different, whereas in the second example the whole plots are chosen such that they are expected to be similar.

The first example consists of a 2^4 factorial design for the qualitative factor and the design for the quantitative factors is constructed using the "top-down"-design in Dehlendorff et al. (2011). The design has ten quantitative factor settings for each whole plot. The four qualitative factors are: anaesthesiologists (2 or 3), porter (3 or 4), recovery beds (6 or 8) and operating days (5 or 4). Operating days is the number of days with elective surgery, i.e., four days implies longer days compared to five days. We treat the factors as qualitative, since the number of levels of the factors is small and hence interpolation may not be reasonable. In Dehlendorff et al. (2011) this data set was analyzed by a generalized additive model (GAMs) (Hastie and Tibshirani, 1990; Wood, 2006). In this paper we however use a constant seed, which makes the output deterministic, and hence the performance of the

GAM models is updated.

The second example has 20 qualitative factor settings, which were chosen from an initial design such that their predicted CVaR waiting time would be short. For each whole plot 20 quantitative factor settings are tested and the design was constructed by the "top-down" method as for the first example. These 20 qualitative factor combinations have 6 active factors: porters (4-5), operating days (4-5), operating rooms (3-4), recovery beds (9-12), cleaning teams (2-4) and increase in elective patient volume (0-5). The second example was in Dehlendorff et al. (2010a) also analyzed by GAM, where it was shown that these settings give better and more robust performance compared to the existing setup of the unit. The model is however in this paper kept in a deterministic operating mode through a constant seed.

6.1 Performance

In Table 4 the 2-stage Kriging model's performance in terms of predicting the CVaR waiting time in the first example at $16 \times 5 = 80$ new sites is summarized and compared with the methods discussed previously. As mentioned earlier in this example the 16 whole plots are generated to perform differently in terms of the CVaR waiting time. It can be seen that the 2-stage model is performing better than the GAM model and the other Kriging models.

In the second example $20 \times 5 = 100$ new quantitative factor settings are used as test cases. The prediction performance for the 2-stage model is better than the other Kriging models, but not as good as the GAM model. This indicates that the Kriging models tend to overfit the data. In both examples

Model	Correlation structure	Example 1	Example 2
Kriging	$\alpha_{ik} = \theta_c$	16.72	1.78
	$\alpha_{ik} = g(\mu_i, \sigma_i)$	9.71	2.00
	2-stage	**9.04**	1.68
	HRM	11.93	1.83
	ZQZ	9.54	1.75
GAM		12.08	**1.27**

Table 4: Performance of models measured in MSPE

it is seen that the 2-stage model is the best Kriging model followed by the model by ZQZ.

6.2 Discussion

The 2-stage model proposed in this paper is seen to give good fits for the examples considered. The model by HRM was seen to give poorer fits compared to the 2-stage model. This may be explained by the additional information contained in the m Kriging models fitted for each whole plot. The model by ZQZ is seen to perform better than the model by HRM, but not as good as the 2-stage model. This may be explained by the complexity of this model compared to the 2-stage model. In the example with 20 different qualitative factor settings the correlation model proposed by ZQZ consists of 209 parameters, whereas the 2-stage procedure uses 16 parameters (eight for the quantitative factor and eight for the whole plots).

It should be noted that the model by ZQZ is a more general model, however for simpler applications it may result in overfitting. The overfit is primarily related to the potentially huge number of parameters used for the correlation matrix corresponding to the correlation between whole plots. However, in

cases with negative correlation between whole plots the model by ZQZ may perform better. More data may also improve the model, but the number of experiments is often limited and hence a trade-off between meta-model accuracy and simulation time should be taken into account.

7 Conclusion

In this article we introduced a Kriging model for computer experiments with qualitative and quantitative factors. Estimation of the model parameters consisted of two stages and was shown to perform better compared to other Kriging models. However, the resulting model is more complex and has more parameters compared to some of the other Kriging models considered in this article, which implies that the time needed for fitting the model may be of concern. The recently proposed model by Hung et al. (2009) was shown to give a poorer fit even with the same number of parameters. Moreover, it was seen that for the examples considered the flexible model proposed by Zhou et al. (2010) did not perform as well as the 2-stage model. This model was furthermore seen to require many parameters, which makes the estimation slow and may require more data.

Typically a single run in a computer or simulation model can take long time, which implies that the added time for estimating a more complex model is less of a concern compared to using extra runs. The proposed method is more efficient than analyzing the qualitative factor combination separately and hence requires fewer experiments. Moreover, the proposed 2-stage procedure

can easily be implemented since it only involves a series of simple Kriging models, which are commonly used in practice.

References

Alexander, S., T. Coleman, and Y. Li (2006). Minimizing cvar and var for a portfolio of derivatives. *Journal of Banking and Finance 30*(2), 583–605.

Ankenman, B., B. L. Nelson, and S. Jeremy (2008). Stochastic kriging for simulation metamodeling. In *Proceedings of the 2008 Winter Simulation Conference*, pp. 362–370.

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2008). Designing simulation experiments with controllable and uncontrollable factors. In *Proceedings of the 2008 Winter Simulation Conference, Miami, FL, 2008.*

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2010a). Analysis of computer experiments with multiple noise sources. *Quality and Reliability Engineering International 26*(2), 137–46. DOI: 10.1002/qre.1035.

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2011). Designing simulation experiments with controllable and uncontrollable factors for applications in health care. *Journal of Royal Statistical Society: Series C 60*(1). DOI: 10.1111/j.1467-9876.2010.00724.x.

Dehlendorff, C., M. Kulahci, S. Merser, and K. K. Andersen (2010b). Conditional value at risk as a measure for waiting time in simulations of hospital units. *Quality Technology and Quantitative Management 7*(3), 321–336.

Dellino, G., J. Kleijnen, and C. Meloni (2009). Robust optimization in simulation: Taguchi and Krige combined. Working paper: `http://center.uvt.nl/staff/kleijnen/RO_Krige.pdf` (July 28th 2010).

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hung, Y., V. Roshan Joseph, and S. N. Melkote (2009). Design and analysis of computer experiments with branching and nested factors. *Technometrics 51*(4), 354–365.

Johnson, R. T., D. C. Montgomery, B. Jones, and J. W. Fowler (2008). Comparing designs for computer simulation experiments. In *Proceedings of the 2008 Winter Simulation Conference*, pp. 463–470.

Kibzun, A. and E. Kuznetsov (2003). Comparison of var and cvar criteria. *Automation and Remote Control 64*(7), 153–164.

Kibzun, A. I. and E. A. Kuznetsov (2006). Analysis of criteria var and cvar. *Journal of Banking & Finance 30*(2), 779–796.

Kleijnen, J. P. (2008a). *Design and Analysis of Simulation Experiments*. Springer.

Kleijnen, J. P. (2008b). Design of experiments: Overview. In *Proceedings of the 2008 Winter Simulation Conference*, pp. 479–488.

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research 192*(3), 707–716.

Krahl, D. (2002). The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*, pp. 205–213.

Lophaven, S., H. Nielsen, and J. Søndergaard (2002a). Aspects of the matlab toolbox dace. Technical Report IMM-REP-2002-13, Informatics and Mathematical Modelling, Technical University of Denmark. `http://www.imm.dtu.dk/~hbn/publ/TR0213.ps`.

Lophaven, S., H. Nielsen, and J. Søndergaard (2002b). Dace - a matlab kriging toolbox version 2.0. Technical Report IMM-REP-2002-12, Informatics and Mathematical Modelling, Technical University of Denmark. `http://www.imm.dtu.dk/~hbn/publ/TR0212.ps`.

Martin, J. D. and T. W. Simpson (2005). Use of kriging models to approximate deterministic computer models. *AIAA Journal 43*(4), 853–863.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology 58*(8), 1246–1266.

Qian, P. Z. G., H. Wu, and C. J. Wu (2008). Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics 50*(3), 383–396.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments. *Statistical Science 4*(4), 409–423.

Santner, T. J., B. J. Williams, and W. I. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer.

van Beers, W. C. and J. P. Kleijnen (2008). Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research 186*(3), 1099–1113.

Wood, S. (2006). *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC.

Zhou, Q., P. Z. Qian, and S. Zhou (2010). A simple approach to emulation for computer models with qualitative and quantitative factors. Working paper: `http://www.stat.wisc.edu/~zhiguang/qpqq2.pdf`.

Bibliography

Ankenman, B. E., B. L. Nelson, and J. Staum (2010). Stochastic kriging for simulation metamodeling. *Operations Research 58*(2), 371–382.

Banks, J., J. S. Carson II, B. L. Nelson, and D. M. Nicol (2005). *Discrete-Event System Simulation* (Fourth ed.). Pearson Education, Inc.

Bettonvil, B. and J. P. Kleijnen (1997). Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research 96*(1), 180–194.

Bielen, F. and N. Demoulin (2007). Waiting time influence on the satisfaction-loyalty relationship in services. *Managing Service Quality 17*(2), 174–193.

Brailsford, S. C. (2007). Tutorial: Advances and challenges in healthcare simulation modelling. In *Proceedings of the 2007 Winter Simulation Conference*, pp. 1436–1448.

Bursztyn, D. and D. Steinberg (2006). *Screening: Methods for Experimentation in Industry, Drug Discovery and Genetics*, Chapter Screening Experiments for Dispersion Effects, pp. 21–47. Springer New York. Editors: A. Dean and S. Lewis.

Chang, P., B. Williams, T. Santner, W. Notz, and D. Bartel (1999). Robust optimization of total joint replacements incorporating environmental variables. *Transactions of the ASME. Journal of Biomechanical Engineering 121*(3), 304–310.

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2008). Designing simulation experiments with controllable and uncontrollable factors. In *Proceedings of the 2008 Winter Simulation Conference, Miami, FL, 2008.*

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2010a). Analysis of computer experiments with multiple noise sources. *Quality and Reliability Engineering International 26*(2), 137–46. DOI: 10.1002/qre.1035.

Dehlendorff, C., M. Kulahci, and K. K. Andersen (2011). Designing simulation experiments with controllable and uncontrollable factors for applications in health care. *Journal of Royal Statistical Society: Series C 60*(1). DOI: 10.1111/j.1467-9876.2010.00724.x.

Dehlendorff, C., M. Kulahci, S. Merser, and K. K. Andersen (2010b). Conditional value at risk as a measure for waiting time in simulations of hospital units. *Quality Technology and Quantitative Management 7*(3), 321–336.

Dellino, G., J. Kleijnen, and C. Meloni (2009). Robust optimization in simulation: Taguchi and Krige combined. Working paper: `http://center.uvt.nl/staff/kleijnen/RO_Krige.pdf` (July 28th 2010).

Donohue, J. (1995). The use of variance reduction techniques in the estimation of simulation metamodels. In *Simulation Conference Proceedings, 1995. Winter*, pp. 194 –200.

Fang, K.-T., R. Li, and A. Sudjianto (2006). *Design and Modeling for Computer Experiments*. Chapman & Hall/CRC.

Fang, K.-T. and C.-X. Ma (2001). Wrap-around l2-discrepancy of random sampling, latin hypercube and uniform designs. *Journal of Complexity 17*(4), 608–624.

Ferrin, D. M. and D. L. McBroom (2007). Maximizing hospital financial impact and emergency department throughput with simulation. In *Proceedings of the 2007 Winter Simulation Conference*, pp. 1566–1573.

Gross, D. and C. M. Harris (1998). *Fundemental of Queueing Theory* (Third ed.). Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC.

Hung, Y., V. Roshan Joseph, and S. N. Melkote (2009). Design and analysis of computer experiments with branching and nested factors. *Technometrics 51*(4), 354–365.

Johnson, M. E., L. M. Moore, and D. Ylvisaker (1990). Minimax and maxmin distance design. *Journal of Statistical Planning and Inference 26*(2), 131–148.

Kibzun, A. and E. Kuznetsov (2003). Comparison of var and cvar criteria. *Automation and Remote Control 64*(7), 153–164.

Kibzun, A. I. and E. A. Kuznetsov (2006). Analysis of criteria var and cvar. *Journal of Banking & Finance 30*(2), 779–796.

Kleijnen, J. and W. van Beers (2004). Application-driven sequential designs for simulation experiments: Kriging meta-modeling. *Journal of the Operational Research Society 55*, 876–883.

Kleijnen, J. P. (2008). *Design and Analysis of Simulation Experiments*. Springer.

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: A review. *European Journal of Operational Research 192*(3), 707–716.

Krahl, D. (2002). The extend simulation environment. In *Proceedings of the 2002 Winter Simulation Conference*, pp. 205–213.

Lant, T., M. Jehn, O. M. Araz, and J. W. Fowler (2008). Simulation pandemic influenza preparedness plans for a public university: A hierarchical system dynamics approach. In S. Mason, R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. Fowler (Eds.), *Proceedings of the 2008 Winter Simulation Conference, Miami*, pp. 1305–1313.

Law, Awerill M. and Kelton, W. David (2000). *Simulation Modeling and Analysis* (3rd ed.). McGraw-Hill.

Li, R. and A. Sudjianto (2005). Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics 47*(2), 111–120.

Lophaven, S., H. Nielsen, and J. Søndergaard (2002a). Aspects of the matlab toolbox dace. Technical Report IMM-REP-2002-13, Informatics and Mathematical Modelling, Technical University of Denmark. `http://www.imm.dtu.dk/~hbn/publ/TR0213.ps`.

Lophaven, S., H. Nielsen, and J. Søndergaard (2002b). Dace - a matlab kriging toolbox version 2.0. Technical Report IMM-REP-2002-12, Informatics and Mathematical Modelling, Technical University of Denmark. `http://www.imm.dtu.dk/~hbn/publ/TR0212.ps`.

Martin, J. D. and T. W. Simpson (2005). Use of kriging models to approximate deterministic computer models. *AIAA Journal 43*(4), 853–863.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology 58*(8), 1246–1266.

McKay, M., R. Beckman, and W. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*(2), 239–245.

Mellor, G. R., C. S. Currie, E. L. Corbett, and R. C. Cheng (2007). Targeted strategies for tubercoulosis in areas of high hiv prevalence: A simulation study. In *Proceedings of the 2007 Winter Simulation Conference*, pp. 1487–1493.

Montgomery, D. C. (2009). *Design and Analysis of Experiments* (7th ed.). John Wiley and Sons, Inc.

Myers, R., D. Montgomery, and C. Anderson-Cook (2009). *Response surface methodology: process and product optimization using designed experiments* (3rd ed.). Wiley, New York.

Qian, P. Z. G., M. Ai, and C. F. J. Wu (2009a). Construction of nested space-filling designs. *The Annals of Statistics 37*(6A), 3616–3643. DOI: 10.1214/09-AOS690.

Qian, P. Z. G., B. Tang, and C. J. Wu (2009b). Nested space-filling designs for computer experiments with two levels of accuracy. *Statistica Sinica 19*, 287–300.

Qian, P. Z. G. and C. F. J. Wu (2009). Sliced space-filling designs. *Biometrika 96*(4), 945–956.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Sacks, J., S. B. Schiller, and W. J. Welch (1989a). Designs for computer experiments. *Technometrics 31*(1), 41–47.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989b). Design and analysis of computer experiments. *Statistical Science 4*(4), 409–423.

Santner, T. J., B. J. Williams, and W. I. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer.

Schruben, L. W. and B. H. Margolin (1978). Pseudorandom number assignment in statistically designed simulation and distribution sampling experiments. *Journal of the American Statistical Association 73*(363), 504–520.

Taguchi, G. (1987). *System of experimental design, volumes 1 and 2*. UNIPUB/Krauss International, White Plains, New York.

van Beers, W. and J. Kleijnen (2003). Kriging for interpolation in random simulation. *Journal of the Operational Research Society 54*, 255–262.

van Beers, W. and J. Kleijnen (2004). Kriging interpolation in simulation. a survey. In R. Ingalls, M. Rosetti, J. Smith, and B. Peters (Eds.), *Proceedings of the 2004 Winter Simulation Conference*, pp. 113–121.

van Beers, W. C. and J. P. Kleijnen (2008). Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research 186*(3), 1099–1113.

Wood, S. (2006). *Generalized Additive Models - An Introduction with R*. Chapman & Hall/CRC.

Zhou, Q., P. Z. Qian, and S. Zhou (2010). A simple approach to emulation for computer models with qualitative and quantitative factors. Working paper: `http://www.stat.wisc.edu/~zhiguang/qpqq2.pdf`.